Project Proposal (Assignment 3)

Ojas Patwardhan

Email- patwardhan.o@husky.neu.edu

Saurabh Vaidya

Email- vaidya.saur@husky.neu.edu

Virender Singh

Email- singh.vir@husky.neu.edu

March 2020

1 Title

Finding the factors that impact the rank of top ten songs in spotify and predict whether a song will be able to make it to the top ten list.

2 Objective and significance

2.1 Describe what the goal of the project is, why is it important, and your motivation for doing it

- Companies working in the field of music technologies like The Echo Nest, ChartMetric, and Next Big Sound have been using data analytics to help artists and labels predict and track a song's success for almost a decade. This problem is widely known as Hit song Science in the domain of Music Information Retrieval. The Content-based music information retrieval tasks can be solved in a two step process : features are extracted from music audio signals, and are then used as input to a regressor or classifier. Such an approach is useful in finding the class of a song i.e. whether the song will be a top ranking song and if so to predict the accurate rankings for the same.

- A model for hit song prediction will be very helpful in the music industry to identify emerging trends and the factors which are making songs popular and potential artists or songs before they are marketed to the public.

3 Background

3.1 Introduce all important concepts and background information

- This work is based on a hypothesis that Songs that hit the top ranks share some common features among them. We attempt to find the relations between different features of the song and their ranks. This concept is known as Music Information Retrieval or MIR. Various attributes like lyrics of the songs, danceability, tones etc. can be taken into contribution. Some custom made features like Singer's popularity or Music director's past tracks can be also considered. This whole work can be divided into a three block pipeline. The first is called feature learning which will rely on techniques like visualization and clustering to understand the data better. The second step is classification which given a song can predict whether the song will be ranked in top ten or not. Finally we have regression technique to assign some scores to the song which can help us predict the ranking of the song.

- The methods which we will be using for these purposes are Naive Bayes Classification, Neural Networks, Linear/ Logistic Regression and Decision Trees. Naive Bayes is a generative approach to prediction and can be used for both classification and regression. Generative approaches, learn the joint probability distribution $p(x, y) = p(x|y)p(y)$. Naive Bayes makes the assumption that the features are conditionally independent.

- In Linear regression, we try to learn the relationship between features and target by assuming an underlying linear relationship between them. Basically, the target function is modeled as a linear combination of features and parameters i.e $\sum_{j=0}^{d} w_j x_j$.

- Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. The logistic or sigmoid function is given by the formula $P(Y = 1|x, w) = (1 + e^{-t})^{-1}$.

- Neural networks refer to systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input, so the network generates the best possible result without needing to redesign the output criteria. A decision tree is a largely used non-parametric effective machine learning modeling technique for regression and classification problems. To find solutions a decision tree makes sequential, hierarchical decision about the outcomes variable based on the predictor data.

3.2 Search the literature and describe previous work on this problem

- In [1] authors have collected a dataset of approximately 4000 hit and non hit songs and extracted each song audio features using spotify web API. They predicted which songs will become top 100 billboard hot songs with approximately 75% accuracy. The most successful algorithms were Logistic Regression and a neural network with one hidden layer. Data was collected using [3] which is a research project stemming from a New York Times article that used the Spotify audio features to illustrate the similarity of summer songs. Yu-et al[4] in their paper do it exclusively for hit song prediction for pop music. Authors model a ranking system for pop music and use a convolutional neural network for the same. They train a multi-objective CNN model with Euclidean loss and pairwise ranking loss to learn from audio the relative ranking relations among songs. Researchers [5] have used K-means clustering to find groupings of the songs that are hit and those which are not. They make use of feature learning to learn the features from the data itself. This technique is very effective even with the simple clustering algorithms such as K-means. In [2] authors address the problem of dance hit song prediction to understand what actually makes a song hit.The database has songs spanning from the period 1985-2013. They could successfully design a model that could predict whether the song will be in the top 10 or not.

- Another group of researchers used Support Vector Machines (SVM) to predict top 10 Dance Hits [4]. By narrowing the scope of the study to only dance music, researchers were able to present a more focused work. Another study attempted to classify songs based on lyric content [6]. While they successfully classified many hits, they also returned many false positives and concluded that analyzing lyrics is an ineffective approach to this problem. The work done in [7] tries to learn a classifier that predicts if the author would like the song or not. The positive dataset is all the songs from the author's spotify playlist and negative data is a playlist of the songs the author doesn't like. Several classifications algorithms are implemented and its concluded that decision trees gave the highest accuracy for the given data. Later the author builds a logistic regression model to find out the coefficients of features with highest values to reason why the author likes the songs. Herremans et al. (2014)[8] describe their hit song predictor that uses various audio features and temporal features. They train their models of decision trees, Naive Bayes, SVM, logistic regression and rules based classifier with a slightly imbalanced distribution of positive and negative samples. The ROC AUC is the highest for logistic regression. .

3.3 If there exists previous work on the problem, describe what makes your work distinct or particularly interesting

- Previous works like [1] and [3] have focused on classification and regression techniques for finding out the relevant class of music or they belong and whether they make it to the top chart or not. Our work is mostly focused on finding the reasons a particular song gets hit in its timeline and how a song released in future will perform given these factors. We are also making a rank predictor for the song which will be able to find the rank of the song in the chart.
- They [7] have a similar problem like ours, but the problem is very specific to predicting author's liked and disliked songs. Thus the dataset is all songs from the author's playlist

4 Proposed approach

4.1 Data Gathering

- We aim to collect a dataset of approximately 10000 songs from Spotify and billboard API . The songs that are in the top ten will be labelled as positives and other songs which do not make it to top ten will be labelled as negative. The features of the songs will be extracted using spotify Feature extraction web API. The extracted features by spotify API for each collected song will be preprocessed and converted to numerical form. The data will then have many feature columns, one target binary column indicating whether the song is in top ten or not and one rank column indicating the top rank achieved by the song.

4.2 Proposed method and Implementation

- Our task is divided into three sub-tasks. First is to estimate the features and their impact on the ranking of a song and define an ordering from most impacting feature to least impacting feature. We will take into consideration the correlation of each feature with the target feature itself. Using the visualization technique and dimensionality reduction, we will be able to make inference about this relation.

- Second task is to classify a given song whether it will be in the top ten or not. The model trained this way should be powerful enough to predict a song it has never seen before that whether it will be in the top ten or not. Here we will be implementing various classification models like Random Forest, artificial neural network (ANN) etc. and compare the performance of each model to detect the class of the song. The main classifiers we would try first would be logistic regression, if data is linearly separable, Naive bayes and Neural Networks for non-linear data and Decision Trees for its ability to work well with imbalanced datasets. If these algorithms don't give satisfactory results then we would think of SVM or using ensemble learning.

- However before learning any classifiers, it is important to note that this is an imbalanced label data classification problem. Therefore, we would first try to oversample the positive labels/undersample negative labels to make it a balanced classification problem, or we would use weighted cost function that penalises misclassified hit song more than misclassified non hit song[12].

- And finally we will be predicting the rank of a particular song. This will be done through regression technique where we try to find the points scored by a particular song. These points will then decide the rank of a song. One can say that higher the points, better the rank and lower the points, worse the rank. This can be done by finding the polynomial regression form for the data.

4.3 Evaluation strategy

- For the first part of our project, we are planning to use precision, recall which was also used in [9] and F-1 score for evaluation purposes. The reason behind using precision and recall is that our data set is imbalanced, i.e we have more negative cases and only a small number of positive cases.

- The formulas for F-1 score, precision and recall are given below:
  F-1 score $= 2\times\frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$ and $precision = \frac{\text{true positives}}{\text{true positives + false positives}}$ and $recall = \frac{\text{true positives}}{\text{true positives + false negatives}}$

- In addition to the above metrics as used in [11], we will also use confusion matrix and precision-recall plot.

- For the second part, we will use MSE and R^2 for evaluation of our model.

- $MSE = \frac{1}{n}\Sigma_{i=1}^{N}(\hat{y_i} - y_i)^2$ and $R^2 = \frac{\Sigma_{i=1}^{N}(y_i-\hat{y_i})^2}{\Sigma_{i=1}^{N}(\bar{y_i}-y_i)^2}$ where $y_i - \hat{y_i}$ is actual - predicted and $\bar{y_i} - y_i$ is mean of predicted - actual

4.4 Describe expected outcomes with a fall back on option in case the initial idea fails

- The expected outcome is firstly building a classifier which will output whether or not a song will be in the top 10 list. After building the above classifier we will also use linear regression to predict the rank of a song.
- In case the initial idea fails, we will still build a classifier to distinguish between songs which will be in top 10 and those that will not be in the list.
- If we get a poor accuracy during classification, we will use the weights of the individual features to deduce which features are having the most impact in getting a song in the top 10 and which ones are playing very little part in doing so. We can also add new features such as lyrics of those songs similar to [10].
- If ranking / classification are not accurate enough due to how the data inherently is, then we would segregate the data points into their genres and build a classifier for the highest 2 dataset genres.
- If the classifier does not have good metrics for correctly classifying hit songs due to imbalanced dataset, we could simply sample lesser non-hit songs to make it a balanced dataset and run our classification algorithms.

5 Individual tasks

- Ojas Patwardhan : Ojas will be working on generating features of each song by using the Spotify and Billboard web API's to create the data. Using techniques like correlation matrix and univariate selection, he will be performing feature selection. Also, if time permits, he will try to make a web app where a user can enter a song's name and in the background our ML model will predict the ranking of the song and whether it will make it to the top ten.
- Saurabh Vaidya : Saurabh will be performing preprocessing on the dataset and will be building classifier models and comparing their performance with actual rankings. He will be evaluating models on the basis of evaluation methods like precision, recall mentioned above.
- Virender Singh : Virender will focus on predicting the rank of a given song. This will be done using regression models which will predict some score of a song. Based on this score, the ranking of a song will be decided. A higher scoring song will be better ranked and vice-versa.

6 References

References

[1] http://cs229.stanford.edu/proj2018/report/16.pdf

[2] https://www.tandfonline.com/doi/pdf/10.1080/09298215.2014.881888?needAccess=true

[3] Guo, A. Python API for Billboard Data. Github.com. Retrieved from: https://pypi.org/project/billboard.py/

[4] https://arxiv.org/abs/1710.10814

[5] https://pdfs.semanticscholar.org/8c03/0a736512456e9fd8d53763cbfcac0c014ab3.pdf

[6] https://ismir2014.ismir.net/LBD/LBD12.pdf

[7] https://opendatascience.com/a-machine-learning-deep-dive-into-my-spotify-data/

[8] Dorien Herremans, David Martens Kenneth Sörensen (2014) Dance Hit Song Prediction, Journal of New Music Research, 43:3, 291-302, DOI: 10.1080/09298215.2014.881888

[9] https://techxplore.com/news/2019-09-spotify-songs.html

[10] http://www.maikaisogawa.com/wp-content/uploads/2019/07/CS_221_Predicting_the_Popularity_of_Top_100_Billboard_Songs.pdf

[11] https://pdfs.semanticscholar.org/e6cc/edb50d2c2b01bca108cb090943e86fb58135.pdf

[12] https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28 (techniques for imbalanced classification)