MAULANA AZAD

NATIONAL INSTITUTE OF TECHNOLOGY

BHOPAL INDIA, 462003

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Hit Song Predictor

Minor Project Report

Semester V

Submitted by:

Vivek Jain	171112305
Siddharth More	171112218
Arsude Akash Nagnath	171112290
Pranshu Kumar	171112297

Under the Guidance of

Dr. Praveen Kaushik

Session: 2019-2020

MAULANA AZAD

NATIONAL INSTITUTE OF TECHNOLOGY

BHOPAL INDIA, 462003

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the project report carried out on Hit Song Predictor by the 3rd year students:

Vivek Jain	171112305
Siddharth More	171112218
Arsude Akash Nagnath	171112290
Pranshu Kumar	171112297

Have successfully completed their project in partial fulfilment of their Degree in Bachelor of Technology in Computer Science and Engineering.

Dr.Praveen Kaushik
(Minor Project Mentor)

DECLARATION

We, hereby declare that the following report which is being presented in the Minor Project Documentation Entitled as Hit Song Predictor is an authentic documentation of our own original work and to the best

Proposed Work and Methodology. of our knowledge. The following project and its report, in part or whole, has not been presented or submitted by us for any purpose in any other institute or organization. Any contribution made to the research by others, with whom we have worked at Maulana Azad National Institute of Technology, Bhopal or elsewhere, is explicitly acknowledged in the report.

Vivek Jain	171112305
Siddharth More	171112218
Arsude Akash Nagnath	171112290
Pranshu Kumar	171112297

ACKNOWLEDGEMENT

With due respect, we express our deep sense of gratitude to our respected guide and coordinator Dr. Praveen Kaushik, for his valuable help and guidance. We are thankful for the encouragement that he has given us in completing this project successfully.

It is imperative for us to mention the fact that the report of minor project could not have been accomplished without the periodic suggestions and advice of our project guide Dr. Praveen Kaushik and project coordinators Dr. Dhirendra Pratap Singh and Dr. Jaytrilok Choudhary.

We are also grateful to our respected director Dr. N. S. Raghuwanshi for permitting us to utilize all the necessary facilities of the college.

We are also thankful to all the other faculty, staff members and laboratory attendants of our department for their kind cooperation and help. Last but certainly not the least; we would like to express our deep appreciation towards our family members and batch mates for providing the much needed support and encouragement.

ABSTRACT

Exploring the possibility of predicting hit songs is both interesting from a scientific point of view and something that could be beneficial to the music industry.

In this project, we are approaching the Hit Song Science problem, which aims to predict which songs will become Billboard Hot 100 hits. We have a dataset of approximately 10,000 hit and non-hit songs. The features and attributes of songs can be collected from spotify web API. We will predict whether the song will make it to Billboard's hot 100 using different Machine Learning Models such as Logistic Regression, Neural Networks , Decision Tree, Random Forest,SVM and make a comparison of accuracies obtained by applying above models.

TABLE OF CONTENTS

1.Introduction

The Billboard Hot 100 Chart [1] remains one of the definitive ways to measure the success of a popular song. The Billboard Hot 100 is the music industry standard record chart in the United States for songs, published weekly by Billboard magazine.

The increasing amount of digital music available online today and the progression of technology have changed the way we listen to and consume music.

We will use machine learning techniques to predict whether or not a song will become a Billboard Hot 100 hit, based on its audio features. The input to each algorithm is a series of audio features of a track. We use the algorithm to output a binary prediction of whether or not the song will feature on the Billboard Hot 100. Not only will it help determine how best to produce songs to maximize their potential for becoming a hit, it could also help decide which songs could give the greatest return for investment on advertising and publicity. Furthermore, it would help artists and music labels determine which songs are unlikely to become Billboard Hot 100 hits.

The theory behind this science is that hit songs are related in the sense that they have a set of features that make them appealing to a majority of people. These features can be processed using machine learning where the aim is to find patterns in the data. If a pattern can be found, this could be used to predict if a song will become a hit.

There are certain characteristics for hit songs, what are the largest influencers on a song's success, and even old songs can predict the popularity of new songs. To make these predictions, we will use the Million Song Dataset provided by Columbia, Spotify's API, and machine learning prediction models. We will predict and compare the accuracy of different classifiers such as Logistic Regression,Neural Networks,SVM,Random Forest,Decision Tree.

2.Literature and review

Prediction tasks are common in machine learning. Thus, there is already a lot of literature on predicting the popularity of songs, or predicting hit songs. These projects ask questions similar to ours: are there certain characteristics for hit songs? What has the largest influence on a song's success? Can old songs predict the popularity of new songs? Music is likely a common area of

interest, as it adds another layer of complexity beyond text and natural language alone. This project will utilize Spotify's API to gather metadata on tracks. They also defined the "success" of a song by whether or not it made it onto the Billboard Hot 100 chart. This project used models like logistic regression and SVM as we did, but they also looked at models like KNN, and a decision tree.

The article, "Predicting Hit Songs with Machine Learning," by Minna Reiman and Philippa Ornell was referred.From this article we inferred that the potential problem in using only audio features seems not to be sufficient information for predicting a hit.

This project theorized that hit songs had features in common that made them appealing to a majority of people. They also utilized Spotify's API. The other models used for this project were K-Nearest Neighbors and Gaussian Naive Bayes. Their most accurate model was their Gaussian Naive Bayes model, with 60.17% accuracy. This project explored the changes in certain music features over time. This is further analysis of the data and results that would be great to include in a more rigorous approach of this problem. Compared to projects similar to ours, our main differential is our data set and our balance of text vs audio features. Our data set is well balanced, and it is also modernized to contain songs from a similar age. These are components that may have led to a higher accuracy rate with our models.

In 2008, Pachet and Roy wanted to validate the hypothesis that popularity of songs can be predicted from acoustic or human features. Their research was based on a dataset of 32000 titles and 632 features, which can be seen as a massive number of features to consider. They were not able to develop a good classification model and claimed that the popularity of a song cannot be learnt by using state-of-the-art machine learning (Pachet & Roy 2008)

However, in 2011, a study made by Ni et al. provided a more optimistic result on the problem of predicting music popularity. They had the goal of distinguishing the top 5 from the bottom 10 in a top 40 hit list. Their dataset was based on UK charts during a time period of 50 years. 5947 unique songs were collected from the Official Charts Company (OCC), and the audio features were extracted from The Echo Nest. Their results indicated that it is possible to identify hits.

In 2011, Borg and Hokkanen investigated if they could predict the popularity of a song based on its audio features and Youtube view counts. They draw the conclusion that audio features alone do not seem to be good predictors of what makes a song popular.

Another related research carried out by Herremans et al. in 2014, focuses on classification of dance hit songs. This research also extracted audio features from The Echo Nest. Some of the machine learning algorithms they used where: Decision tree, Naive Bayes, Logistic Regression and Support Vector Machine. This study showed that the popularity of dance hit songs can indeed be predicted from analysing audio features. They concluded that the overall best algorithm was Logistic Regression, with an accuracy of 0.65 and a precision of 83% (Herremans et al. 2014).

3.Gaps Identified.

The review of several research papers and ongoing studies revealed several gaps with regard to the Billboard hit song prediction using machine learning techniques. Machine learning is a popular research and industry tool to approach the HSS question. Researchers have used Convolutional Neural Networks and K-Means Clustering to predict pop hits. Both of these studies were engaging and successful, but focused more heavily on the signal-processing involved in audio analysis. Different types of audio features were used to predict hit songs such as Danceability, Energy, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Loudness, and Tempo. In

our study, we identified that along with these features, some other features have become very predominant in hit song prediction such as artist, genre , song length. We aim to use these features also to predict hit songs along with audio features.

4.Proposed Work and Methodology.

i) Dataset and features

A dataset of 10,000 random songs will be collected from the Million Songs Dataset (MSD) , a free dataset maintained by labROSA at Columbia University and EchoNest.

The audio features of the songs will be extracted from spotify web API.The datasets provided the artist name and song title, as well as other miscellaneous features. To balance the dataset between positive (hits) and negative (non-hits) examples, we removed two thirds of the songs collected from the Billboard Hot 100.

Tracks were labeled 1 or 0: 1 indicating that the song was featured in the Billboard Hot 100 (between 1991-2010) and 0 indicating otherwise. Next, we used the Spotify API to extract audio features for these songs.

The Spotify API provides users with 13 audio features, of which we chose nine for our analysis: **Danceability, Energy, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Loudness, and Tempo**. The first seven features are represented as values between 0 and 1 by Spotify.Loudness is measured in decibels and tempo refers to the speed of the song in beats per minute.

To account for artist recognisability, we define an additional metric: the artist score. Each song will be assigned an artist score of 1-100 based on the number of hits the artist had previously.

ii) Methods and Algorithms Used

To obtain the audio features of the tracks we used Spotify's open Web API. To retrieve the data from the Spotify API a lightweight Python library will be used, called **Spotipy.**

We will use different algorithms such as Logistic Regression, K-Nearest Neighbours, Neural Networks,Gaussian Naive Bayes, and Support Vector Machine.The selected algorithms are a mix of linear and nonlinear models. The models will be then tested on the test data set and their accuracies will be recorded and compared.

The description of various algorithms to be used is as follows:

a) Logistic Regression:

Logistic regression is a mathematical model which can be used to describe the relationship between one or more independent variables and one dependent variable. This model is therefore used for problems where the outcome can be classified in one of two categories. When applying the estimated logistic model to new cases of a test dataset, it provides a prediction of success probability, which is a number between 0 and 1. The logistic regression provides a rule for classifying the test data with a cut-off on the predicted success probability.

b) Support Vector Machines

A Support Vector Machine is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In a two dimensional space this hyperplane is a line dividing a plane in two parts with one class on each side of the line. In a 3 dimensional space, this separator is instead a plane separating the different classes. The Support Vector Machine attempts to find a separating hyperplane with a margin that is as large as possible from the nearest data points.

c) K Nearest Neighbour

K-Nearest Neighbours classification implements learning based on the K nearest neighbours of each query point. This method is a type of non-generalizing learning since it only stores instances of the training data, it does not create an internal model for generalizing the data. The classification is evaluated from a majority vote of the nearest neighbours of each query point, and each point is assigned to the class which is the most common among its neighbours. The choice of the value of k is dependent on the dataset: if k is set to a low value, the point is assigned to a class with only a few neighbours, and if k is a high value, then the effect of noise is suppressed and the classification is of a more general form.

d) Neural Networks

Neural Networks are the biomimicry of human brain.Neural networks are multi-layer networks of neurons that we use to classify things, make predictions, etc.They are inspired by biological neural networks and the current so-called deep neural networks have proven to work quite well. Neural Networks are themselves general function approximations, which is why they can be applied to almost any machine learning problem about learning a complex mapping from the input to the output space.

5.Tools and technologies to be used:

The various tools and technologies to be used are as follows:

i) Python Libraries to implement Machine Learning Models -

- Pandas - pandas is a software library written for the Python programming language for data manipulation and analysis.
- Numpy - NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- Scikit learn - Scikit is an open source Python library that implements a range of machine learning, pre-processing, cross-validation and visualization algorithms using a unified interface.
- Spotipy - *Spotipy* is a Python library for the Spotify Web API. With *Spotipy* we get full access to all of the music data provided by the Spotify platform.

To build the machine learning models we used The Jupyter Notebook . Jupyter Notebook is an open-source web application that can be used to implement statistical modelling, data visualisation, machine learning etc.

6.Conclusion

The Billboard Hot 100 Chart is one of the definitive ways to measure the success of a popular song.This project is relevant to musicians and music labels. Not only will it help determine how best to produce songs to maximize their potential for becoming a hit, it could also help decide which songs could give the greatest return for investment on advertising and publicity. Furthermore, it would help artists and music labels determine which songs are unlikely to become Billboard Hot 100 hits. This will help musicians to review their work and improve the quality of song. In this synopsis, we have introduced a machine learning based model for predicting hit songs based on audio features, artist, genre. This project will also help the music labels to review songs.

7.References

1. https://towardsdatascience.com
2. scikit-learn: machine learning in python scikit-learn 0.17 documentation
3. HITPREDICT: PREDICTING HIT SONGS USING SPOTIFY DATA STANFORD COMPUTER SCIENCE 229: MACHINE LEARNING by Elena Georgieva, Marcella Suta, and Nicholas Burton.