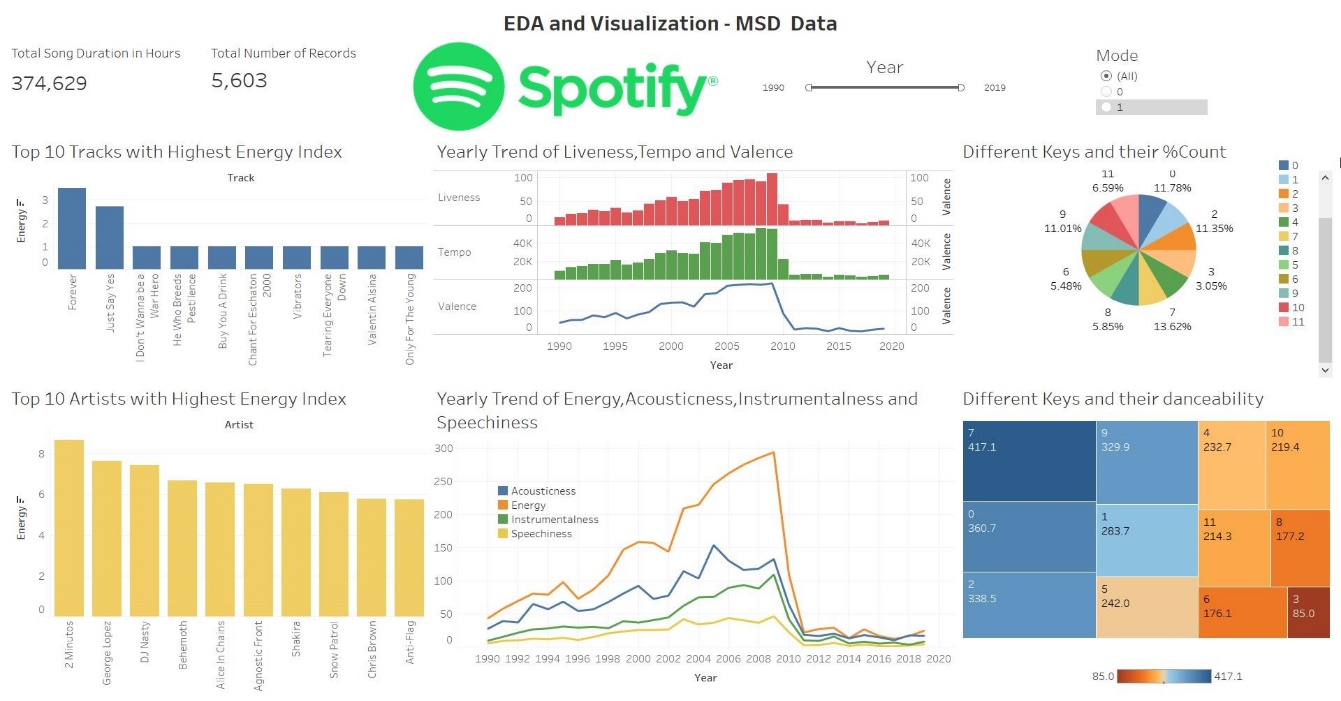
TECHNOCOLABS DATA SCIENTIST INTERNSHIP

**PROJECT REPORT**

**TITLE: Hit Predict-Predicting Billboard Hits Using Spotify Data**

****

**AIM:**

The principle focus of our project is to perform data analysis and train a model using the most popular Machine Learning algorithms like Logistics Regression, Linear and Quadratic Discriminant Analysis Support Vector Machine, Random Forest Classifier, etc in order to analyze the historical data that is present in different data sources like Spotify data, Million Songs Data and Billboard Charts data put all together.

# ABSTRACT:

The Billboard Hot 100 Chart1 remains one of the definitive ways to measure the success of a popular song. We investigated using machine learning techniques to predict which songs will become Billboard Hot 100 Hits.

# INTRODUCTION:

* We extracted different audio features from the Spotify API4 (Table 1).

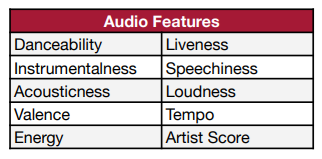
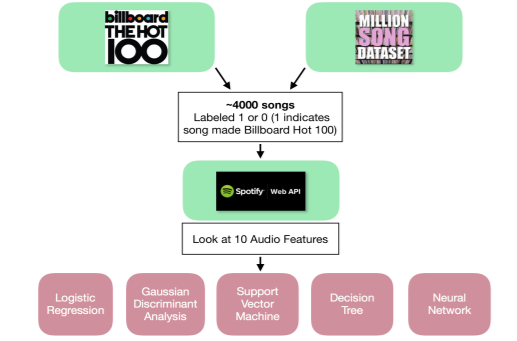


Table 1

* We created the Artist Score metric, assigning a score of 1 to a song if the artist previously had a Billboard hit, and 0 otherwise.



* Data for ~4000 songs was collected from Billboard.com3 and the Million Song Dataset5. Songs were from 1990-2018.
* Songs were labeled 1 or 0 based on Billboard success.
* Audio features for each song were extracted from the Spotify Web API4.
* Different machine-learning algorithms were used to predict a song’s Billboard success.

# OVERVIEW:

* Data Segmentation and Data Cleaning
* Exploratory Data Analysis using python’s data visualization libraries.
* Training the model based on the historical data available.

# DATASET:

# 

* SpotifyID: Unique ID for tracks in Spotify API.
* Track: Track Name.
* Artist: Track performer/singer/band.
* Danceability: Danceability is measured using a mixture of song features such as beat strength, tempo stability, and overall tempo.
* Energy: song energy is the sense of forward motion in music, whatever keeps the listener engaged and listening.
* Key:  the key of a piece is the group of pitches, or scale, that forms the basis of a song.
* Mode: mode, in music, any of several ways of ordering the notes of a scale according to the intervals they form with the tonic, thus providing a theoretical framework for the melody.
* Speechiness: Speechiness detects the presence of spoken words in a track.
* Instrumentalness: This value represents the amount of vocals in the song.
* Liveness: This value describes the probability that the song was recorded with a live audience.
* Valence:  Describes the musical positiveness conveyed by a track.
* Tempo:  tempo is how fast or slow a piece of music is performed.
* Duration\_ms: Track duration in millisecond.
* Loudness: loudness is the subjective perception of sound pressure.
* Is\_hit: identifier if the song/track is Hit or Not. (Target)

# DATA SEGMENTATION AND DATA CLEANING:

* The dataset had few missing values. We dropped those rows.
* We treated the outliers with z-score.
* We checked the distribution of data after outlier treatment
* We tried different plots and combination plots, to check the relationship of different features within themselves and with target variables. (Scatter plot, bar plot, distribution plot, correlation heatmap)
* We performed Hypothesis test with 95% confidence interval, VIF analysis to make sure which features are statistically significant.

# EXPLORATORY DATA ANALYSIS

# Checking Target Variable is imbalanced or not.

# 

# Target variable is not imbalanced.

# Top 10 Energetic Songs

# 

# Top 10 Danceable songs

# 

# Count of Track Based on Keys

# 

# Distribution of song duration

# 

# Distribution of track tempo

# 

# Relationship Plots: 3D Scatter Plots

# 

# Relationship Plots: 2D Scatter Plots

# 

# 

# Overall Relationship plot: correlation heatmap.

# 

# Statistical Significance of Features based on z-test and chi-square test

# 

# 

# VIF Plot: Multicollinearity check

# 

# Conclusion based on EDA:

