

KANWAL SHEHZADI

02-131212-027

DATA SCIENCE

ASSIGNMENT 03

Q1: Using data analysis, song properties can be used to forecast their popularity. The dataset is provided in CSV format which includes all songs that secured a position in the Top 10 of the Billboard Hot 100 Chart between 1990 and 2010.

Additionally, it includes a selection of supplementary songs that did not achieve Top 10 status. The data is sourced from three primary outlets: Wikipedia, Billboard.com, and EchoNest.

* **timesignature** and **timesignature\_confidence** = a variable estimating the time signature of the song, and the confidence in the estimate
* **loudness** = a continuous variable indicating the average amplitude of the audio in decibels
* **tempo** and **tempo\_confidence** = a variable indicating the estimated beats per minute of the song, and the confidence in the estimate
* **key** and **key\_confidence** = a variable with twelve levels indicating the estimated key of the song (C, C#, . . ., B), and the confidence in the estimate
* **energy** = a variable that represents the overall acoustic energy of the song, using a mix of features such as loudness
* **pitch** = a continuous variable that indicates the pitch of the song
* **timbre\_0\_min**, **timbre\_0\_max**, **timbre\_1\_min**, **timbre\_1\_max**, . . . , **timbre\_11\_min**, and **timbre\_11\_max** = variables that indicate the minimum/maximum values over all segments for each of the twelve values in the timbre vector (resulting in 24 continuous variables)
* **Top10** = a binary variable indicating whether or not the song made it to the Top 10 of the Billboard Hot 100 Chart (1 if it was in the top 10, and 0 if it was not)

A screenshot of a computer program

Description automatically generated

1. Use visualization to analyze the trends of popularity of songs for each year with maximum top 10 songs.

A screenshot of a computer

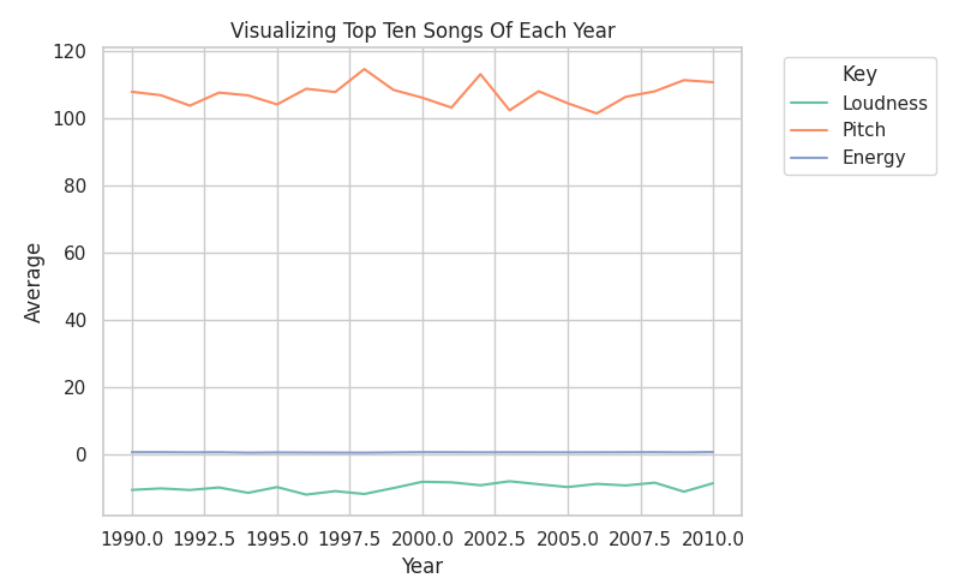
Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

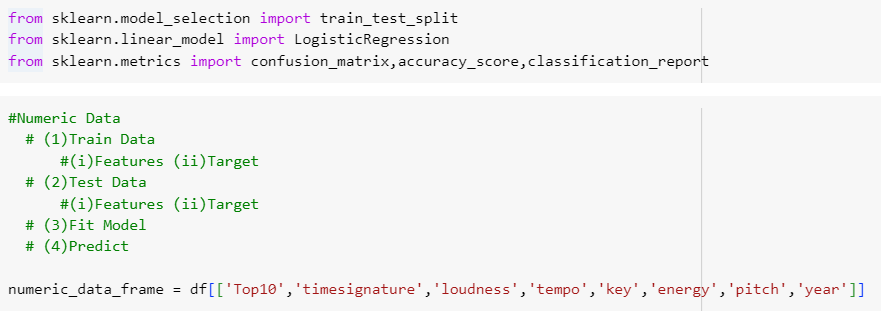
Description automatically generated



(b) Remove the non-numeric variables data from the data frame and find correlation of ‘top 10’ variable with variables: 'timesignature', 'loudness', 'tempo', 'key', 'energy', 'pitch'.



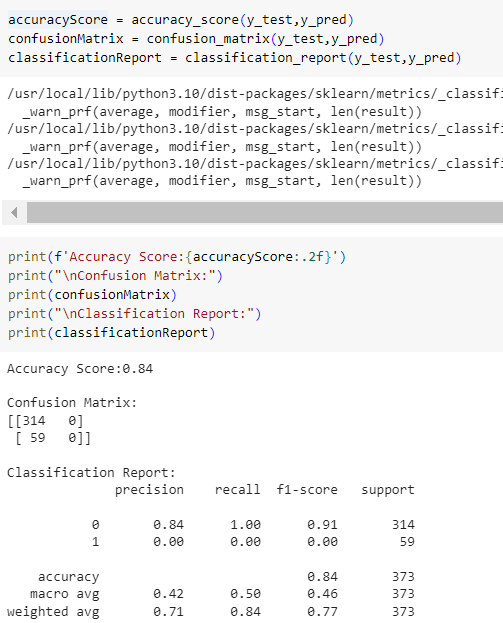
(c) Split data in (c) in training (all data uptill year 2009) and test data(all data year 2010). Train a logistic regression model on the training data and test the model using test data.



A screenshot of a computer program

Description automatically generated

(d) Analyze the performance of model using performance indicators.



**Assignment Notebook Url**

https://colab.research.google.com/drive/1nZ3pA7szOz2dgCFx9aNhmpnSWMQGzks6#scrollTo=IhSaj46enGEL