Project Report

Course- Linear Models

Title – Song Popularity Prediction

Name - Himanshu Vijaysing Pawar (22060641010)

Acknowledgment

We would like to express our sincere gratitude to the various individuals for supporting us throughout our project. Firstly, we would like to thank our project guide, Prof. Priya Deshpande, for her enthusiasm, patience, helpful comments, and unceasing ideas that have helped us at all times during our research and writing this project. Without her support and guidance, this project would not have been possible.

We also wish to express our thanks to Prof. Dr. Sharvari Shukla, Director, Symbiosis Statistical Institute, for always supporting and encouraging us for scientific research and projects.

We are also very grateful to the people who have helped us, directly or indirectly.

Index

Title

ABSTRACT

"Music is the best medicine for the Heart". The music industry consists of several musical parameters. The popularity of the song depends on the song's quality. This project aims to analyze several musical features like energy, loudness, acousticness, and many more. This analysis helps us determine which parameters are playing a vital role in the popularity of the song. If the song is not getting the expected popularity, then which causes are behind that? The analysis of this project is helping the person in the music industry in such a way that; on which things they have to be more concentric.

OBJECTIVES

- To study several musical parameters in song popularity like energy, loudness, acousticness, duration, and many more.
- To study which parameters are significant in the popularity of the song.
- To fit different linear models on the data and check their adequacy.
- To predict the popularity of the song

Data

The data of 200 songs contains popular as well as unpopular songs with parameter values of each song. The data is taken from the Kaggle website which can be accessed through:

https://www.kaggle.com/datasets/amaanansari09/top-100-songs

SOFTWARES USED

Python: Python is a very popular general-purpose interpreted, interactive, object-oriented, and high-level programming language. often used to build websites and software, automate tasks, and conduct data analysis.

Microsoft Excel

Microsoft Excel is a spreadsheet program that offers very helpful tools for data manipulation, analysis, and visualization. For this project, Excel helped us to sort and filter the data according to our requirements and use it later to analyse and give inferences.

Microsoft Word

Microsoft Word is a graphical word-processing software that allows users to type and save documents. It has many tools to make the documents look attractive and professional. For our project, we have used Microsoft Word for the Project Compilation.

Introduction

In this project, we have studied the relationship between the popularity of a song with the impact of its musical parameter values. For example, is there any relationship between high energy or high loudness and the popularity of the song? How is energy gets influenced by loudness? Also, how acousticness in the song is related to energy and loudness? We made an attempt to solve all these queries and tried to come up with accurate predictions.

Variables Explained

i. id – unique song id
ii. name – the name of a song
iii. duration – time duration of a song (in min)
iv. key – particular set of notes or the scale at which is song is performed
v. loudness – loudness intensity in a song measured in decibels (dB)
vi. mode – type of scale (major: 1; minor:0)
vii. speechiness – impact or majority of words in a song (measured over 0 to 1)
viii. acousticness – predominance of acoustic instruments like guitar, piano, or strings i.e. more organic sound (measured over 0 to 1)
ix. liveness – degree of song where it is performed (measured over 0 to 1)
x. valence – overall mood or wing of a song; positive, negative (measured over 0 to 1)
xi. Tempo – speed or pace at which a song is played; measured in beats per minute (BPM)
xii. Danceability – degree of the song to which it is suitable for dance (measured over 0 to 1)

- **Data Cleaning and Editing** –

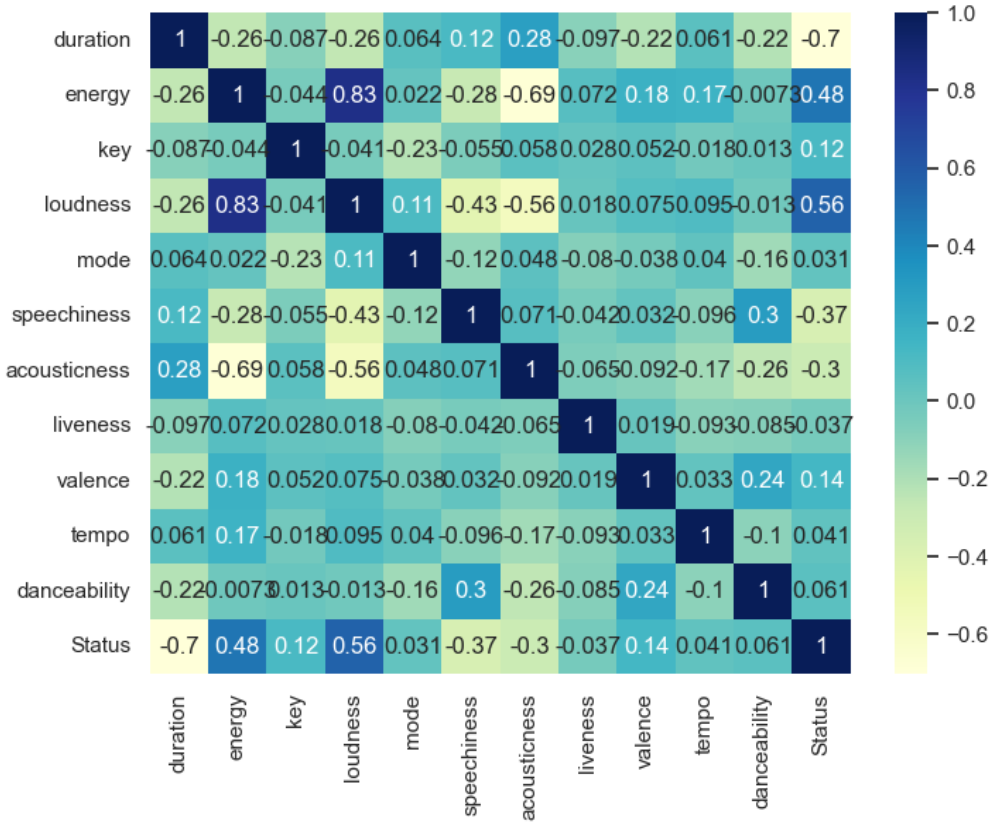
First of all, the data is checked for the null or any misinterpreted values along with the data type of each column. A new binary column is added which contains 1's and 0's denoting a popular song and an unpopular song respectively.

Exploratory Data Analysis

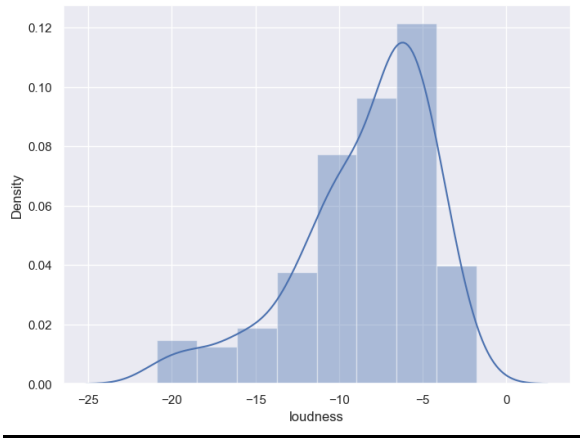
I. Descriptive Statistics – In descriptive statistics, different statistical parameters like mean, standard deviation, minimum and maximum value, median, quartiles, etc. are calculated for every numerical column. This gives an idea about how each variable under consideration is distributed.

II. Correlation Matrix – It gives the correlation between any two variables in the dataset. This matrix is useful to decide the independent variables. The variables that have a significant correlation with the dependent variable or variable under consideration are chosen to fit the respective model.

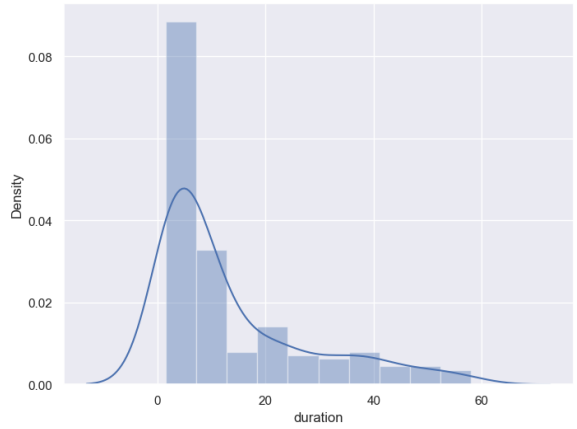
III. Correlation Heatmap – It is a visual representation of a correlation matrix from which we can easily interpret the correlation among these variables.



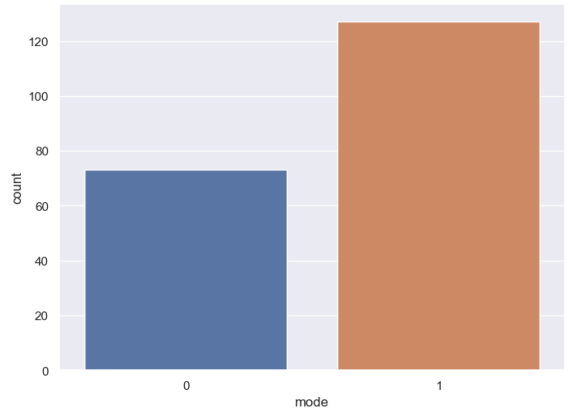
IV. Graphical Visualizations – Different types of graphical visualizations like histograms, box plots, bar graphs, and scatterplots are used to check the distribution of a particular variable, or to compare the effects a variable on the popularity of a song.



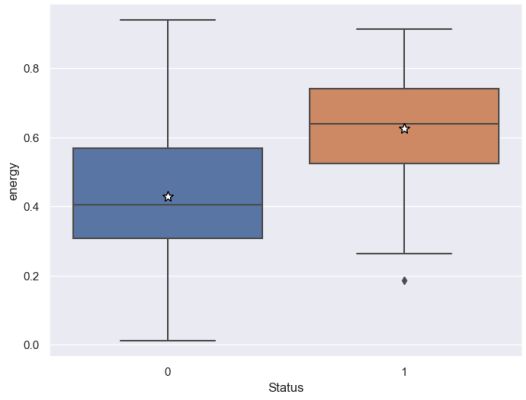
Distribution of loudness: Negatively Skewed



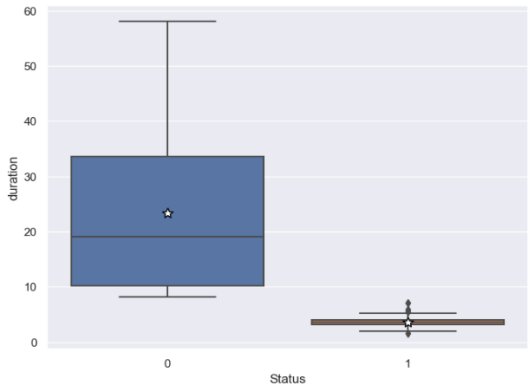
Distribution of loudness: Positively Skewed



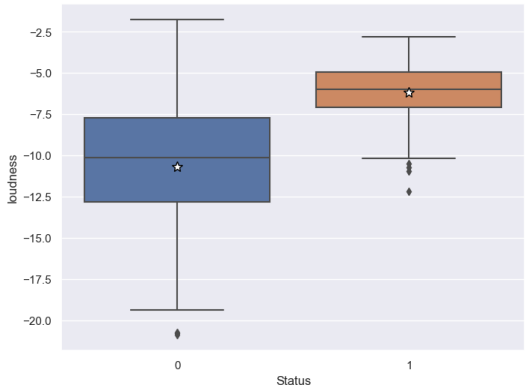
Count plot of mode (0s and 1s)



Status Vs energy

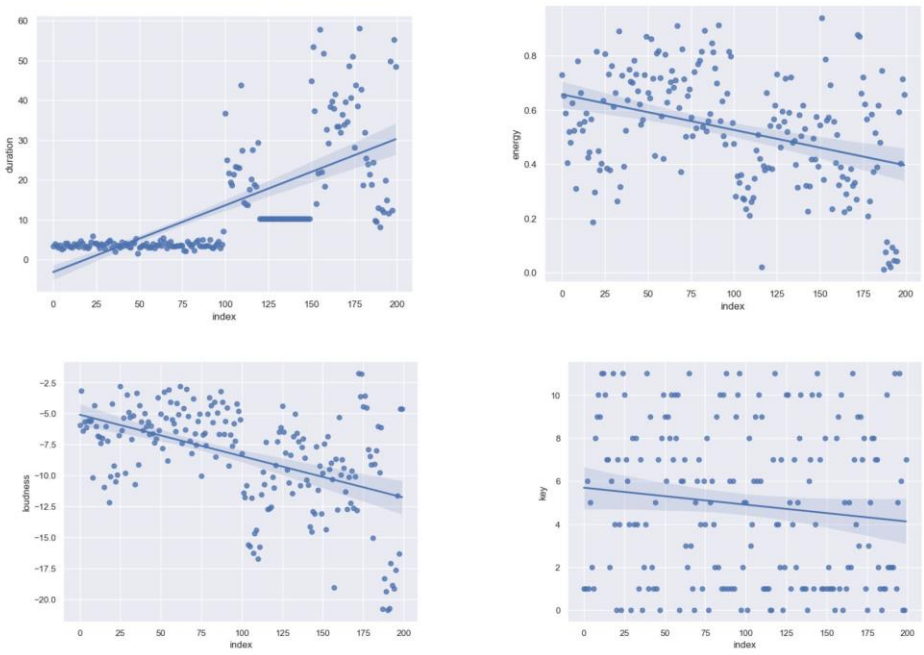


Status Vs duration



Status Vs loudness

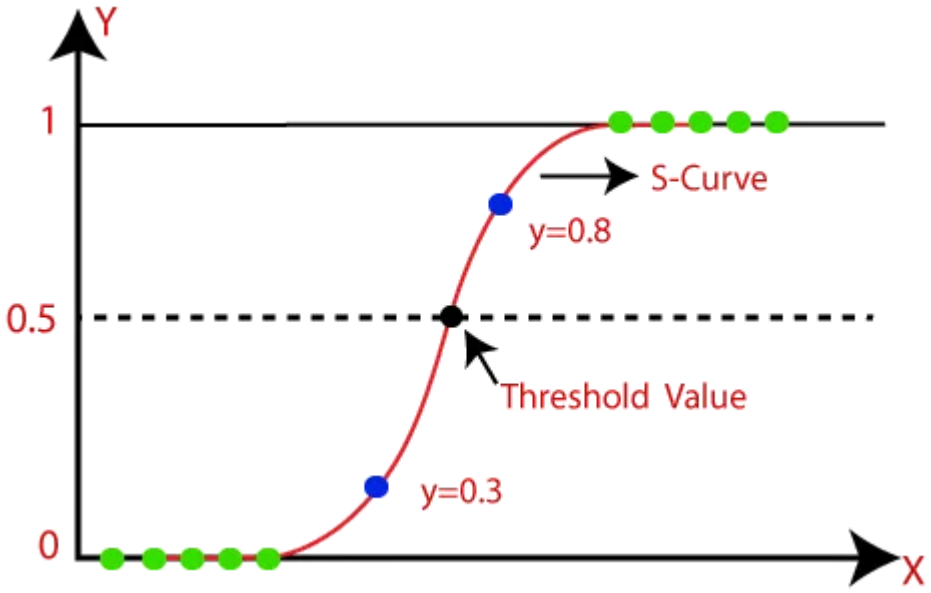
Scatterplots



V. Data Analysis: We used Logistic Regression, Simple Linear Regression, and Multiple Linear Regression in this project. Logistic regression is used to predict whether the song will be popular or not. Simple linear regression and multiple linear regression are also used to predict some musical parameters using other parameters.

Logistic Regression

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.



Graphical Representation of Logistic Regression

- dependent variable - Status
- independent variables - duration, energy, loudness, speechiness, acousticness

Here in logistic regression, the variable under consideration (dependent variable) is whether the song will become popular or not. i.e. Status (0: unpopular; 1: popular). From the scatterplots and correlation matrix, we fixed the independent variables. These are the variables that have a significant impact on the Status (outcome). This is done to get the maximum precision for prediction.

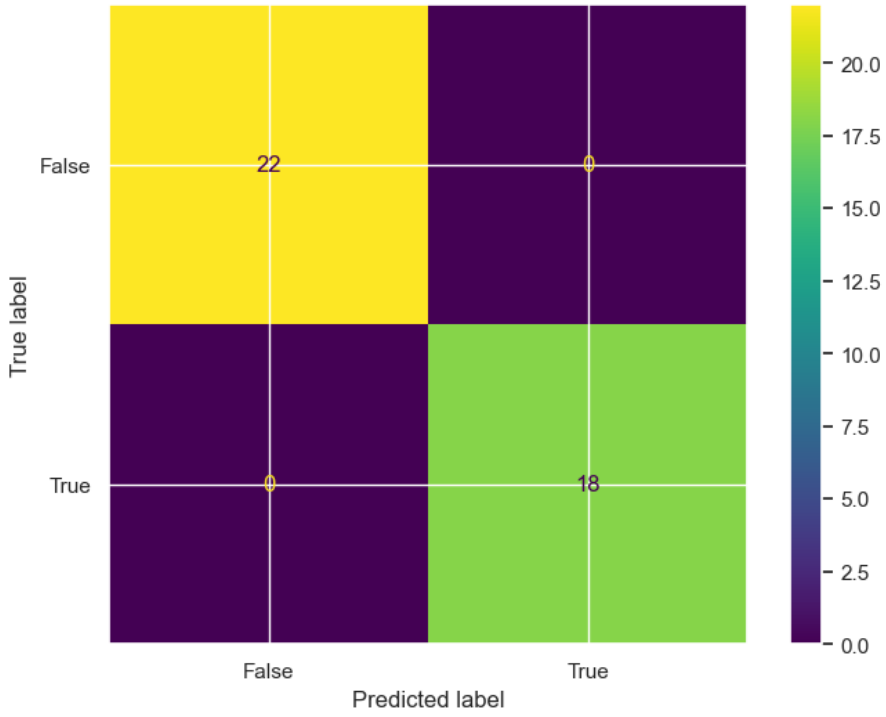
To fit the model, the dataset is split into a training dataset and a testing data set. 80% of the data into the training dataset and the rest 20% in the testing

dataset. The model is trained over the training dataset and it is tested over the testing dataset and the accuracy is measured.

The model is then fitted over the test data set, we got coefficients and intercept values, and predictions are calculated over the rest 20% dataset.

Confusion Matrix –

- Used to check the accuracy of predicted values
- Usually return the correct and incorrect predictions
- Diagonal values are correctly classified values
- Off-diagonal values are incorrectly classified values



Accuracy Score – We got the accuracy score equal to 1 which means all predicted values by this model are correct. The confusion matrix is also showing the same result. The model predicted the song popularity with 100% accuracy.

Model Evolution – For model evolution, we calculated values like MAE, MSE and RMSE

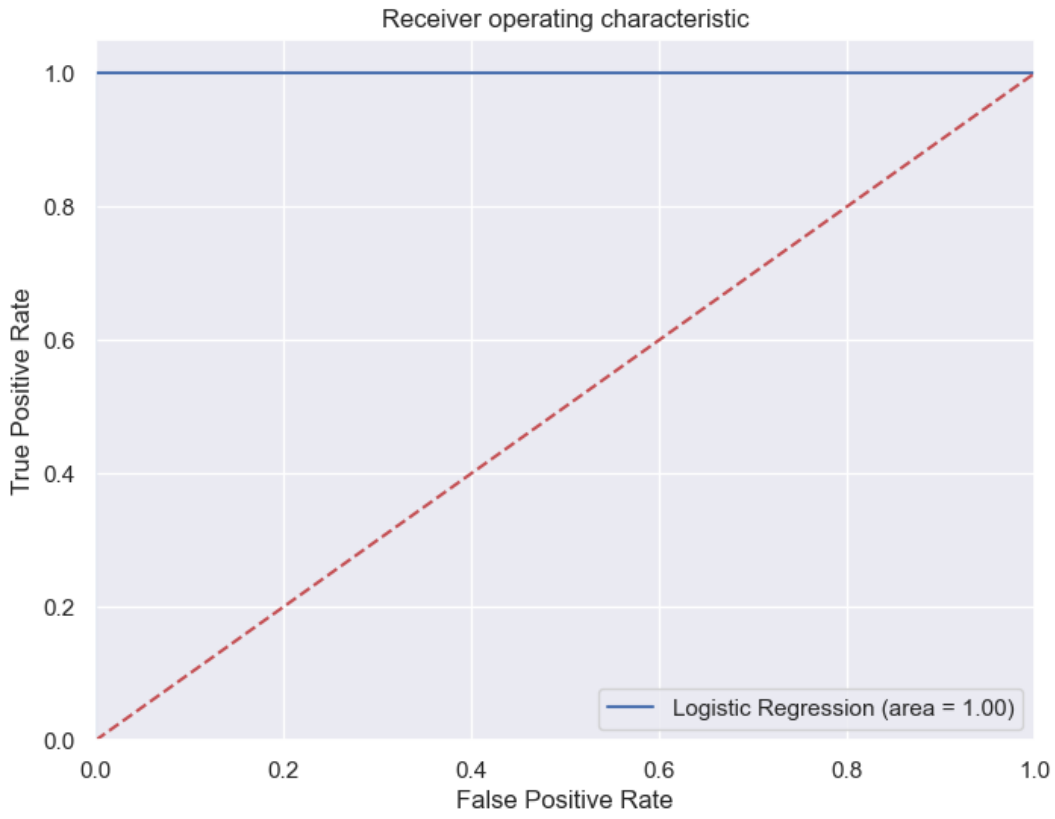
MAE (Mean Absolute Error) = 0.0

MSE (Mean Square Error) = 0.0
RMSE (Root Mean Square Error) = 0.0

All these values are 0 which indicates that the model is good fit.

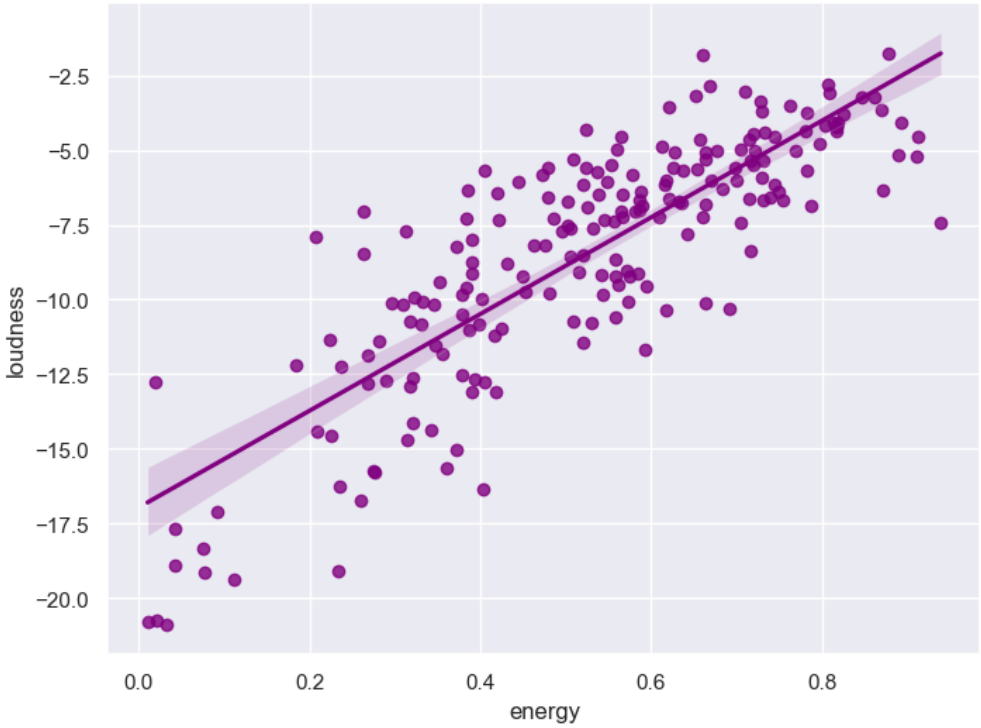
ROC Curve (The receiver operating characteristic (ROC)) –

- The dotted line represents the ROC curve of a purely random classifier
- A good classifier stays as far away from that line as possible (toward the top-left corner).

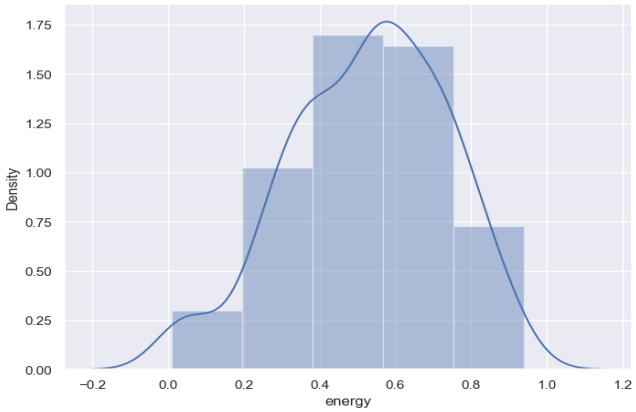


Simple Linear Regression –

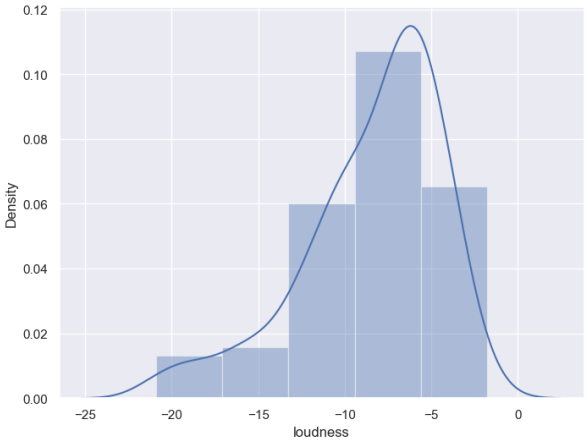
Using the Simple Linear Regression model, we studied the linear relationship between energy and loudness keeping energy as a dependent variable and loudness as an independent variable. The scatterplot of energy versus loudness has also shown a strong positive correlation between them. Energy can't be directly decided hence predicting it using the amount of loudness will definitely help us out in predicting song popularity.



- There is a strong correlation between loudness and energy
- We can predict energy by measuring loudness in a song
- If loudness is high, energy will be high, then the song should be hit
- This can be done using simple linear regression.



Distribution of energy (Normally distributed)



Distribution of loudness (Negatively Skewed)

The dataset is split into training and testing datasets. The model is trained over the training dataset and tested over the testing data set.

After fitting the model, we got the regression equation,

energy = 0.8924 + (0.0430) * loudness

Accuracy Score – The accuracy score came out to be 0.68657; which means 68.66% of values predicted by this Simple Linear Regression model are correct.

Model Evolution –

MAE (Mean Absolute Error) = 0.09914
MSE (Mean Square Error) = 0.01391
RMSE (Root Mean Square Error) = 0.11796

All these values are quite low which indicates that the model is a good fit.

Interpretation – From the above Simple Linear Regression model, if we keep loudness as 0, we will have 0.8924 energy. As soon as we increase (or decrease) the loudness value by a single unit, the energy will increase (or decrease) by 0.0430 units in 0.8924.

Multiple Linear Regression

Multiple Linear Regression is used to study the effect of two or more independent variables on a dependent variable or variable under study. Here, we have studied how the combined effect of energy and loudness reflects acousticness in a song. Acousticness is chosen as the dependent variable because it has a significant negative correlation with energy and loudness which influences the popularity of any song.

The process is similar to Simple Linear Regression as the data is split into a training dataset and a testing dataset and the model is developed over the training dataset and tested over the testing dataset.

After fitting the model, we got the regression equation,

acoustics = 1.0687 + (-1.1931) * energy *+ (0.0083) ** loudness

Accuracy Score – The accuracy score is 0.472512 means only 47.25% of predicted values for acousticness are correct based on energy and loudness.

Model Evolution –

MAE (Mean Absolute Error) = 0.196198
MSE (Mean Square Error) = 0.058951
RMSE (Root Mean Square Error) = 0.242798

All these values are not very low shows that still there is scope for some improvement.

Interpretation – From the obtained Multiple Linear Regression, if we keep energy and loudness value 0, we will get acousticness equal to 1.0687. The unit change in parameters like energy and loudness will affect the acousticness value by -1.1931 and 0.0083 respectively.

Bibliography

i. https://www.kaggle.com/datasets/amaanansari09/top-100-songs
ii. Introduction to Linear Regression Analysis by Douglas C. Montgomery
iii. Programmed Statistics by B L Agrawal

And some more sites and books have helped us gain knowledge to analyze the data better.