Spotify Genre Prediction

Nandini Chaganti
Luddy School of Informatics,
Computing and Engineering
Indiana University Bloomington
Bloomington, Indiana
nchagan@iu.edu

Manya Mallikarjun
Luddy School of Informatics,
Computing and Engineering
Indiana University Bloomington
Bloomington, Indiana
mmallika@iu.edu

Keerthi Reddy Sure
Luddy School of Informatics,
Computing and Engineering
Indiana University Bloomington
Bloomington, Indiana
keersure@iu.edu

Keywords: Genre prediction, One-Hot Encoding, MinMax Scalar, LGBM classifier, Decision Tree Classifier,

I. ABSTRACT

Genres vary from song to song on different properties. Proper classification of these features will help us predict the genre which can be quite useful for customers who are interested in hearing songs from a particular genre. A dataset consisting of these features is collected to perform the classification techniques. Before that, preprocessing steps are employed in order to clean the data. The techniques used to remove null values, negative values and unidentified characters(?). Different modeling techniques[1] are applied to predict the genre and accuracies are computed for each of the models. The results indicate that the LGBM classifier has better performance compared to other classifiers like KNN and ridge.

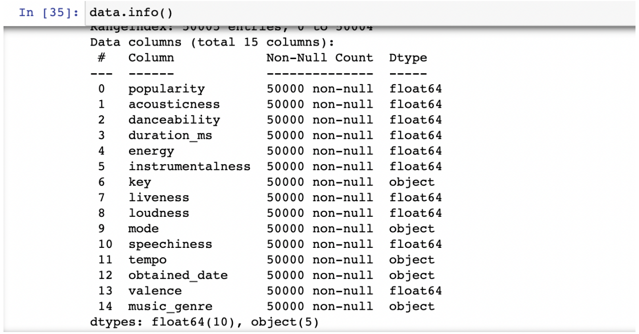
II. INTRODUCTION

Music is a very subjective topic. Nowadays, everyone includes music in their daily lives. There are various kinds of songs. Music tastes vary from person to person. The type of music a person prefers relies on their mood as well as their interests or preferences[6]. But then. If a lot of people enjoy a song, it is unquestionably regarded as a hit because it is popular and is frequently played. Based on the preferences of the listeners, music can be grouped using genre prediction. Spotify can categorize songs from all genres with ease and can even suggest musical subgenres to its customers.

This project aims at predicting the genre for the songs and the genre relies on various factors like danceability, energy, liveliness, tempo, acousticness, etc.

III. DATASET

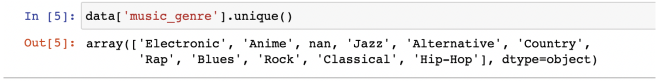
The Dataset contains a varied set of columns, Fig A contains the information about these columns, and the number of rows in the data set is 50000. The columns refer to different properties of the song which differ one song from another.



```
In [35]: data.info()
         RangeIndex: 50005 entries, 0 to 50004
         Data columns (total 15 columns):
          #   Column            Non-Null Count  Dtype
         ---  ------            --------------  -----
          0   popularity        50000 non-null  float64
          1   acousticness      50000 non-null  float64
          2   danceability      50000 non-null  float64
          3   duration_ms       50000 non-null  float64
          4   energy            50000 non-null  float64
          5   instrumentalness  50000 non-null  float64
          6   key               50000 non-null  object
          7   liveness          50000 non-null  float64
          8   loudness          50000 non-null  float64
          9   mode              50000 non-null  object
          10  speechiness       50000 non-null  float64
          11  tempo             50000 non-null  object
          12  obtained_date     50000 non-null  object
          13  valence           50000 non-null  float64
          14  music_genre       50000 non-null  object
         dtypes: float64(10), object(5)
```

Fig A. Analyzing dataset

In the collected dataset, we have 10 types of genres. Which are as follows Electronic, Anime, Jazz, Alternative, Country, Rap, Blues, Rock, Classical, Hip-Hop. All these genres are considered popular genres around the world.



```
In [5]: data['music_genre'].unique()

Out[5]: array(['Electronic', 'Anime', nan, 'Jazz', 'Alternative', 'Country',
               'Rap', 'Blues', 'Rock', 'Classical', 'Hip-Hop'], dtype=object)
```

Fig B. Types of Music Genre

The considered dataset has equal composition of all the genres which is 10%, constituting 100% together, which is the entire dataset.

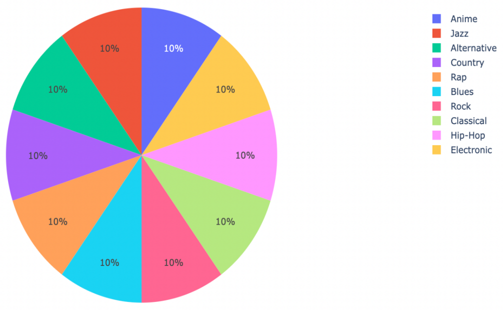
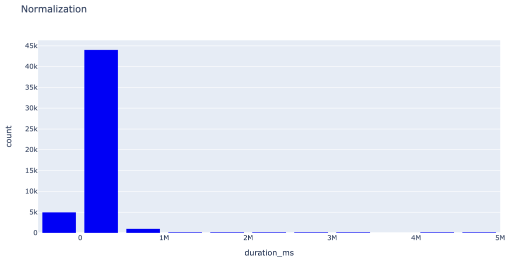


Fig C. Ratio of data for every genre in the dataset

IV. DATA PREPROCESSING

The data is cleaned using different preprocessing techniques which are explained below. These processes helped to obtain better accuracy.

1. *Removing Columns with 0 correlation*:
   track_name and instance_id columns are dropped as they do not have any correlation with other columns, they only have unique values .
2. *Removing NaN and duplicate values*:
   Duplicates and NaN values are dropped using drop_duplicates and dropna.
3. *One-Hot Encoding*:
   It is necessary to transform categorical data into a numerical form and to reclassify model predictions into a categorical form.
   Further, One Hot Encoding is used to convert the categorical data into numeric data.
   With this method, 16 new columns, based on key, mode and obtained_data are generated which are filled with 0's and 1's.
   Key, mode and obtained_data columns are dropped.
4. *Replacing -1 and ? values with the mean value*:
   Duration_ms and tempo columns contain -1 and ? values, which are replaced with the mean values of the same columns.
5. *MinMax Scaler*:
   It is a form of normalization which uses the standard deviation and mean values to map all the available data into the range between min and max values. In most cases, the min and max values are going to be default which are 0 and 1.
   MinMax scalar is performed on the duration_ms column as this column values have a very high range when compared to other columns, so we are using normalization on this column
   Fig D indicates the skewed data for the duration_ms column.



FigD.Normalization graph for duration_,ms

6. *Feature Selection*:

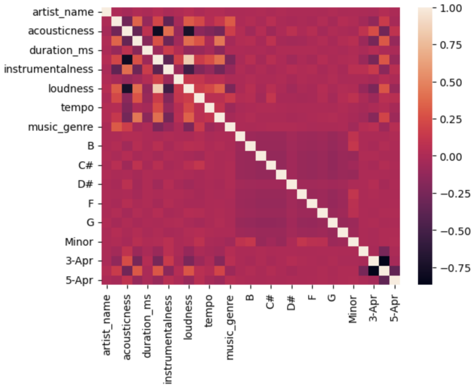


Fig D. Heat Map for columns in the data set

After the observation of the heat map, columns were removed which had a threshold of more than 80 so energy, and 4-Apr columns were dropped as they had a threshold more than 80 when compared with loudness and 3-Apr columns.

The python packages used for this process are Sequential Feature Selector, from mlxtend feature_selection From sklearn linear_model

For feature selection, Linear regression is considered. After successful implementation the code has selected the best 13 features in order to reduce the dimensionality, ignoring the extra columns. These 13 features have created a good balance between the accuracy and efficiency of the model.

The best 13 features that are selected are listed below: popularity, acousticness ,danceability, duration_ms, instrumentalness, liveness, loudness, speechiness, tempo, valence, F, Minor, 3-Apr.

V. MODELS

After preprocessing the music genre data using multiple techniques[4] which were discussed above. The models are trained on 80% of the data and the predictions are made on 20% of the data. All the genres in the data have equal composition. On this data different classification methods are applied which are as follows. The basic definition of these classifiers and the pseudo-code used to obtain the accuracy is as follows.

A.LGBM Classifier

It is a gradient boosting[2] framework that uses tree-based learning algorithms and is regarded as one of the most powerful computation-based algorithms. It is regarded as a processing algorithm with quick speeds. By using this classifier the obtained accuracy is comparatively higher than other classifiers.

B. KNN

The k-nearest neighbor's algorithm, often known as KNN or k-NN, is a supervised learning classifier that makes predictions or classifications about how a single data point will be grouped. Although it can be used to solve classification or regression problems, it is frequently used as a classification technique because it is predicated on the notion that similar points can be found nearby.

C.*Ridge Classifier*

In order to solve the problem, the Ridge Classifier, which is based on the Ridge regression methodology, transforms the label data into the range [-1, 1]. Multiple output regression is used for multiclass data, and the target class is the one with the greatest prediction value.

D. Random Forest

The widely used machine learning technique known as random forest combines the output of multiple decision trees to get a single conclusion. Its adaptability and use have boosted its popularity since it can resolve classification and regression challenges.[3]

E. *Decision Tree*

Models for supervised machine learning include decision tree classifiers. This indicates that they train an algorithm that can make predictions using pre labeled data. Regression issues can also be solved with decision trees.

F.*Support Vector Classifier*

The training examples are plotted in space. There should be an apparent gap between these data points. A straight hyperplane splitting two classes is what is predicted. Maximizing the distance from the hyperplane to the closest data point of either class is the main goal while drawing the hyperplane. As a maximum-margin hyperplane, the depicted hyperplane was referred to.

G.Logistic Regression

Only when a decision threshold is included does logistic regression become a classification approach. The classification problem itself determines the threshold value, which is a crucial component of logistic regression. Based on a given dataset of independent factors, logistic regression calculates the likelihood that an event will occur. Given that the result is a probability, the dependent variable's range is 0 to 1.

H.Naive Bayes

Simple models like the Naive Bayes Classifier are frequently employed in classification issues. The core principles are extremely simple to understand, and the arithmetic that supports them is also quite understandable. However, this model performs surprisingly well in a lot of situations, and it as well as its modifications are utilized to solve a lot of issues.

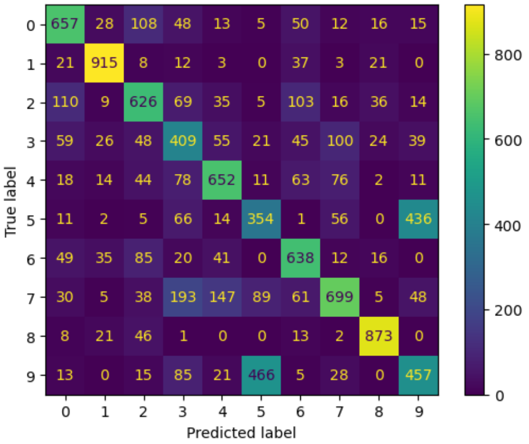
VI. RESULTS

The accuracy is finally obtained after the successful execution of all the defined\classifier models on the given dataset. The classifiers have to predict the correct data genre from the set of ten genres. The classifiers highly vary in the results. The one that stands out with highest accuracy of 62.8% is LGBM Classifier. KNN takes the next place with 51.98% accuracy. Eight different classifiers are used in order to compare and obtain the best accuracy possible. These accuracies are clearly tabularized in table A.

Classifiers	**ROC AUC Score**	**Accuracy**
LGBM Classifier	0.9822	62.8
Logistic Regression	0.7640	29.14
KNN	0.9647	51.98
Support Vector	0.561	17.99
DecisionTree	0.999	43.81
Random Forest	0.884	47.12
Naïve Bayes	0.853	41.81
Ridge	0.465	46.84

Table A. Table with accuracy results of genre prediction

Fig E represents the confusion matrix constructed on top of the predicted data and actual results when the LGBM classifier is used. The Confusion matrix clearly shows the prediction for all the ten genres.



	0	1	2	3	4	5	6	7	8	9
0	657	28	108	48	13	5	50	12	16	15
1	21	915	8	12	3	0	37	3	21	0
2	110	9	626	69	35	5	103	16	36	14
3	59	26	48	409	55	21	45	100	24	39
4	18	14	44	78	652	11	63	76	2	11
5	11	2	5	66	14	354	1	56	0	436
6	49	35	85	20	41	0	638	12	16	0
7	30	5	38	193	147	89	61	699	5	48
8	8	21	46	1	0	0	13	2	873	0
9	13	0	15	85	21	466	5	28	0	457

Fig E. Confusion Matrix for LGBM Classifier

VII. DISCUSSION

From the results, it is clear that the LGBM classifier provides sufficient accuracy as compared to other models like Logistic regression, KNN, Decision tree, support vector, random forest, naive bayes and ridge. It provides an accuracy of 62.8. KNN and random forest provide an accuracy of 51.98 and 47.12 respectively. So, for this particular dataset, using the LGBM classifier, the music genre ‘Anime’ seems to be the best predicted[5] followed by ‘Rock’with this classification technique and the least predicted is ‘Country ’.

VIII. REFERENCES

[1] Luo, Kehan. "Machine Learning Approach for Genre Prediction On Spotify Top Ranking Songs." (2018).

[2] Bang-Dang Pham, Minh-Triet Tran, and Hoang-Long Pham. Hit song prediction based on gradient boosting decision tree. In 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), pages 356–361. IEEE, 2020

[3] Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

[4] Adragna, Robert, and Yuan Hong Bill Sun. "Music Genre Classification." *MIE324 Project Report* (2019).

[5] Huang, Derek A., Arianna A. Serafini, and Eli J. Pugh. "Music Genre Classification." *CS229 Stanford* (2018).

[6] Dawson Jr, Christopher E., et al. "Spotify: You have a Hit!." *SMU Data Science Review* 5.3 (2021): 9