**Finding the genre - Multi Class Classification 12.11.24**

**Made by: Ortal Lasry & Or Cohen Raviv**

**Business question:**

**Background :** Spotify dataset from kaggle, it has 114 track\_genres and the main features are 14 numeric (popularity, duration\_ms, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, time-signature) and one is bool which is called explicit.

The original database has a total of 20 columns and 114,000 rows.

Track\_genre -is the target variable were found with 40,900 duplicates (16,641 unique track\_id and 24,259 redundant) and so we kept only the first occurrence.

After preliminary EDA and data cleansing **87867** rows were left. There is significant imbalance in the target variable (=class imbalance) which is track\_genre. Hence, we will later on control the balancing when fitting model.

From Figure A below we can draw two main conclusions:

a. Some music genres have more track\_id entries than others, making the data unbalanced.

b. There are too many music genres for the human eye to easily interpret patterns.

Additionally, some genres overlap in their characteristics with broader genres.

For example, 'rock', 'rock-n-roll', and 'alt-rock' are different subgenres that fall under the broader category of rock. Hence, we decided to group them to 10 classes (see Figure B).

Figure A. Frequencies of track\_geners

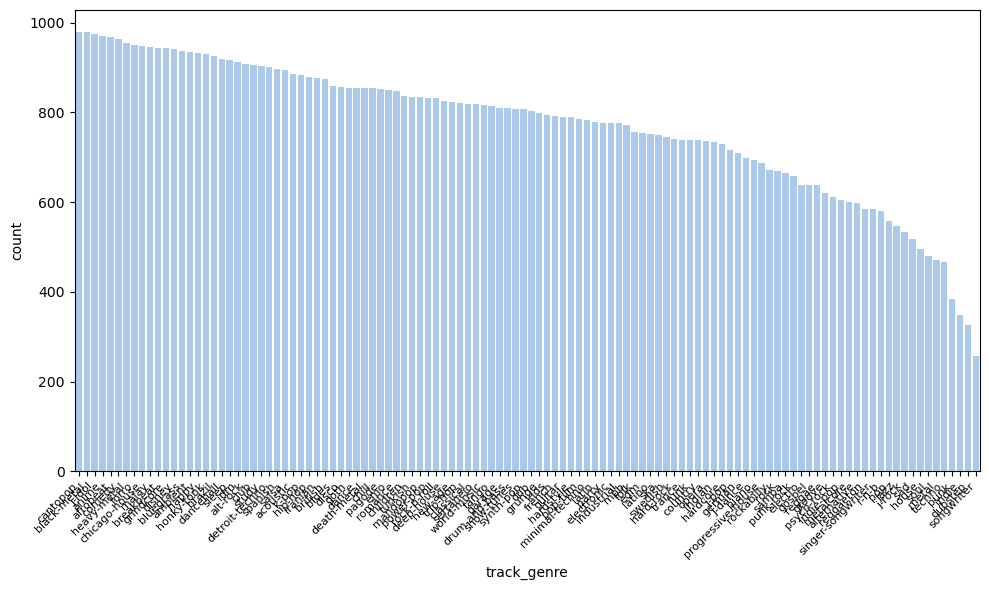
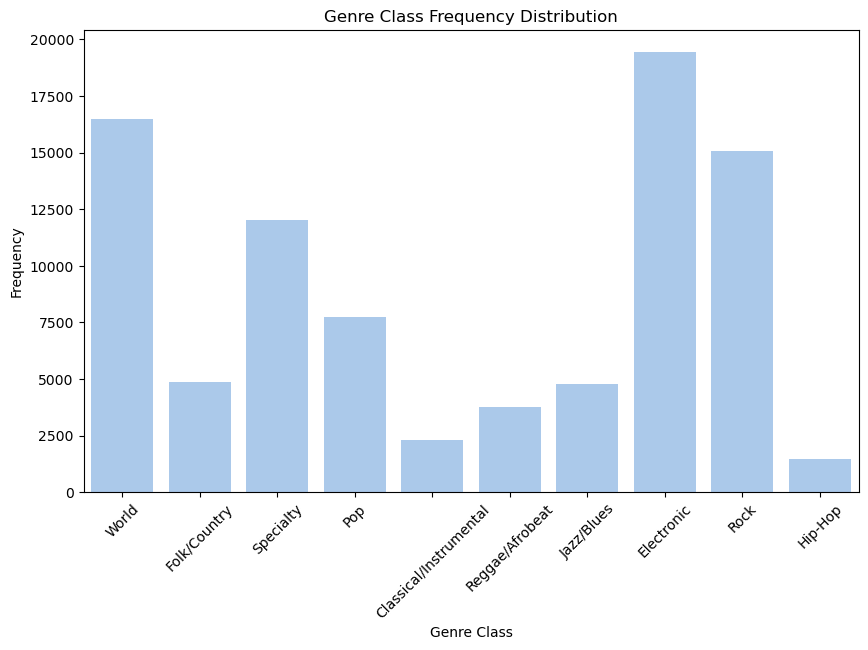
****

Figure B. Frequencies of gener\_class

****

**Feature Engineering:**

**Descriptive statistics of the Features:**

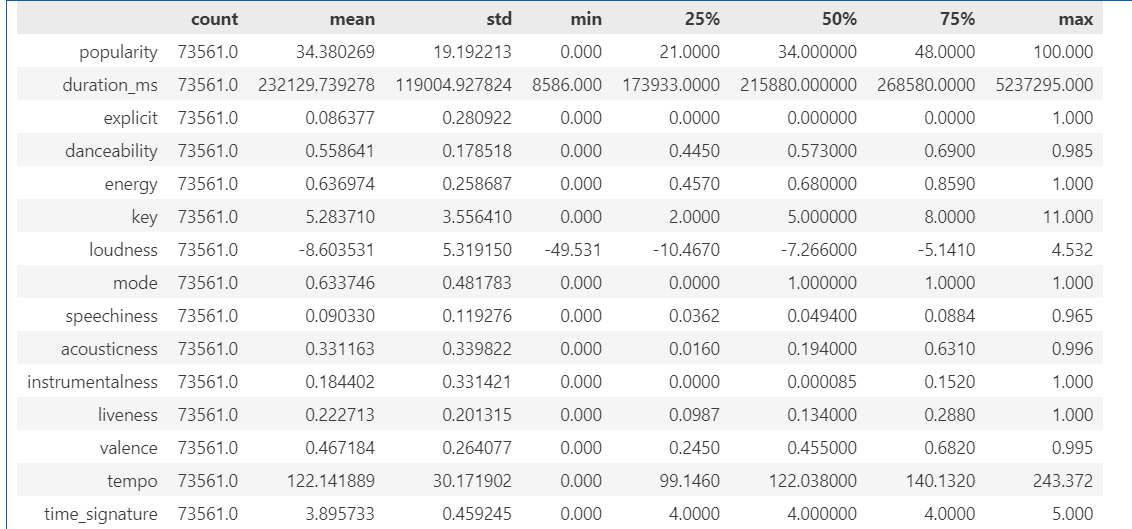
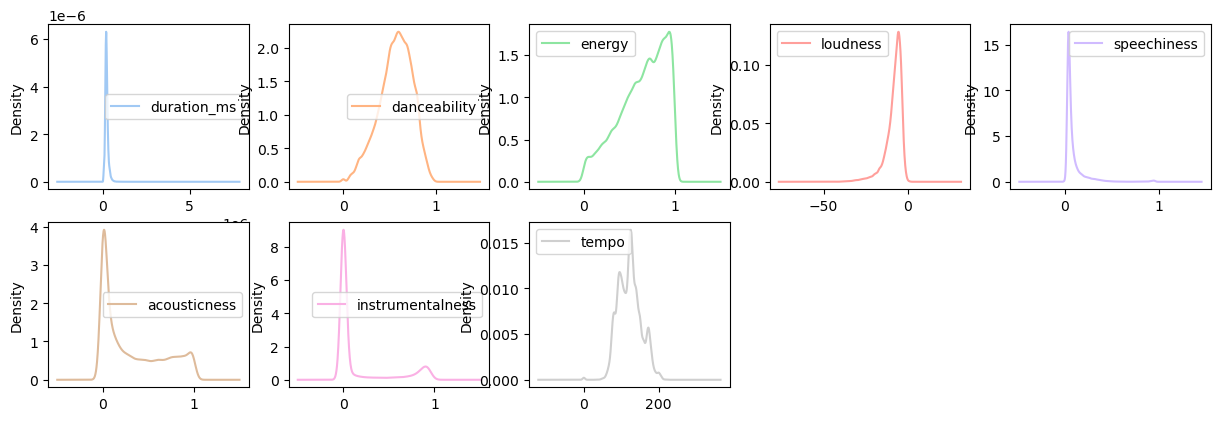
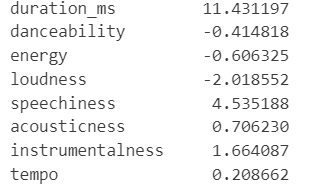
****

Figure C. Density plots for quantitative features



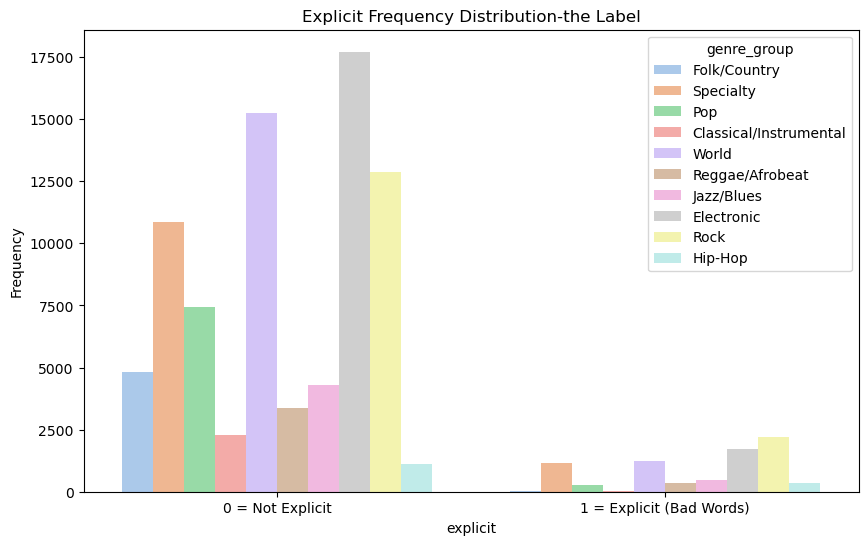
Measuring skewness according to Fisher's skewness:



Next, we investigated the features distributions based on the descriptive statistic table and the figures above, we found a right-skewed distribution for the following variables: duration\_ms, speechiness, acousticness and instrumentalnes, and a left-skewed distribution for loudness(for example, 75% of the cases are below -0.5, whereas the max value of loudness is 4.5).

Bar Chart for the Categorical Feature: 'Explicit' (Presence of Explicit Language).

As expected, some of the genres have higher frequencies of explicit content then others: for example, Hip Hop and Death Metal show the highest frequencies of explicit content, while Classical music has none.

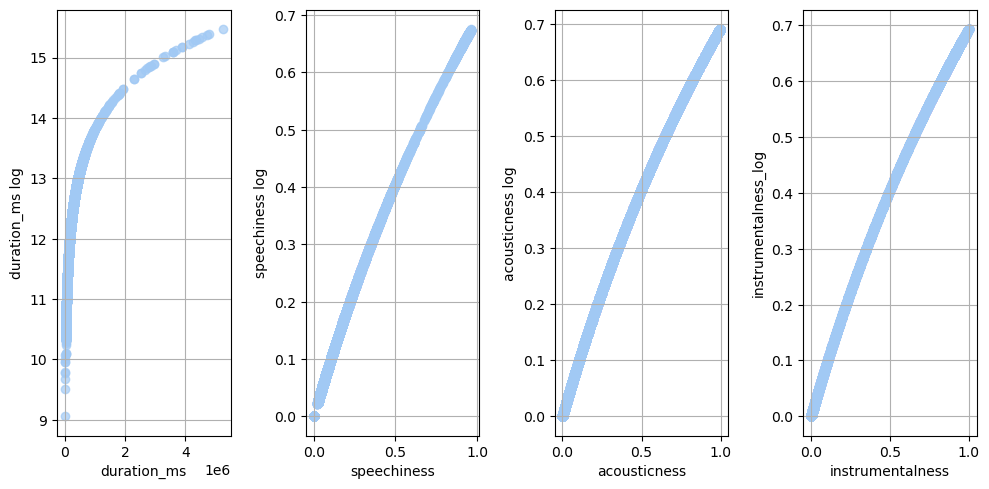


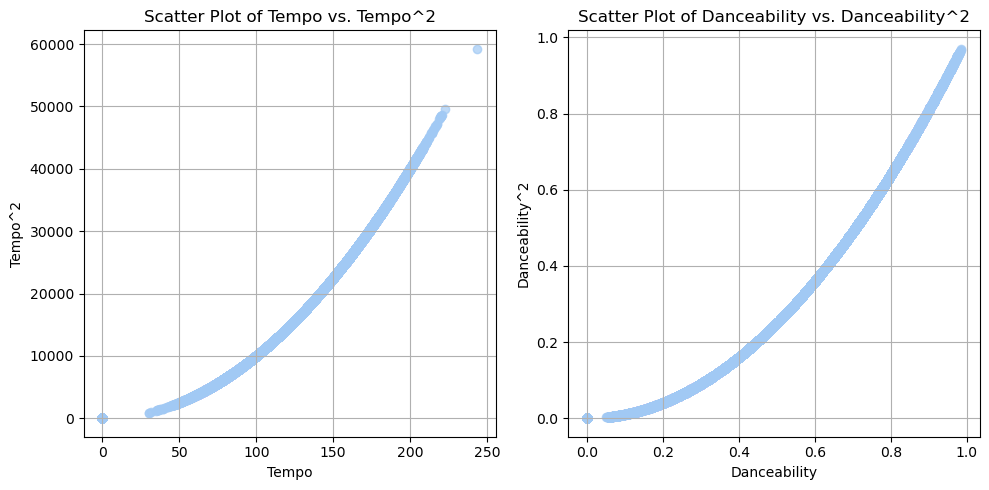
**Feature engineering**

Based on the features distributions above, we found several patterns that we would like to address, in order to capture the feature characteristics better, and to increase the correlation between the features and the classes of the music genres, as well as the level of the model accuracy later on. In order to do so, we decided to run the following transformations: log and polynomial transformation and to generate interaction variables for several features.

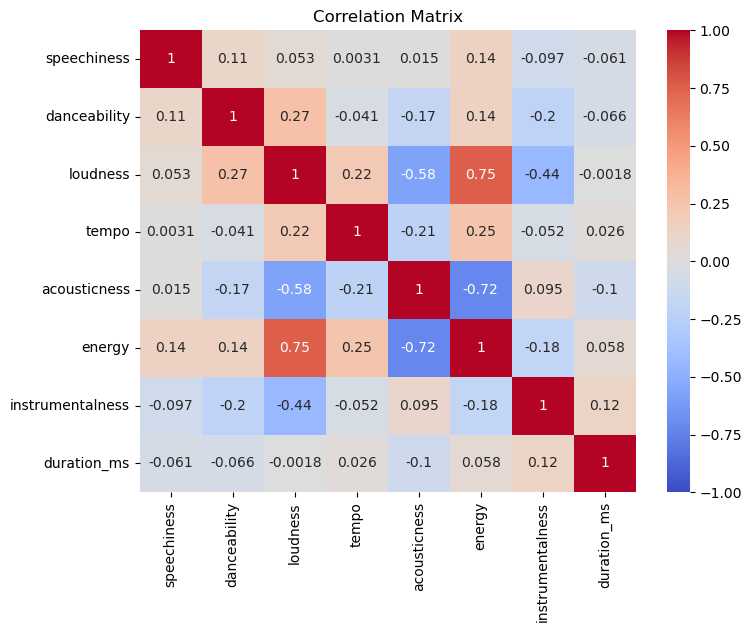
We first apply a log transformation to features that are heavily right-skewed, with most values near zero and a long right tail (e.g., duration\_ms, speechiness, acousticness, instrumentalness). Later on in our models, we will compare the performance of the original variables against their log-transformed counterparts to examine whether the log transformation improves the accuracy of the ML models.

Additionally, since tempo and danceability display non-linear distributions, we will generate polynomial features for these variables. Similarly, in the machine learning models, we will compare the original variables to the polynomial features to assess any accuracy improvements. The distribution below shows the differences between the original and transformed features.



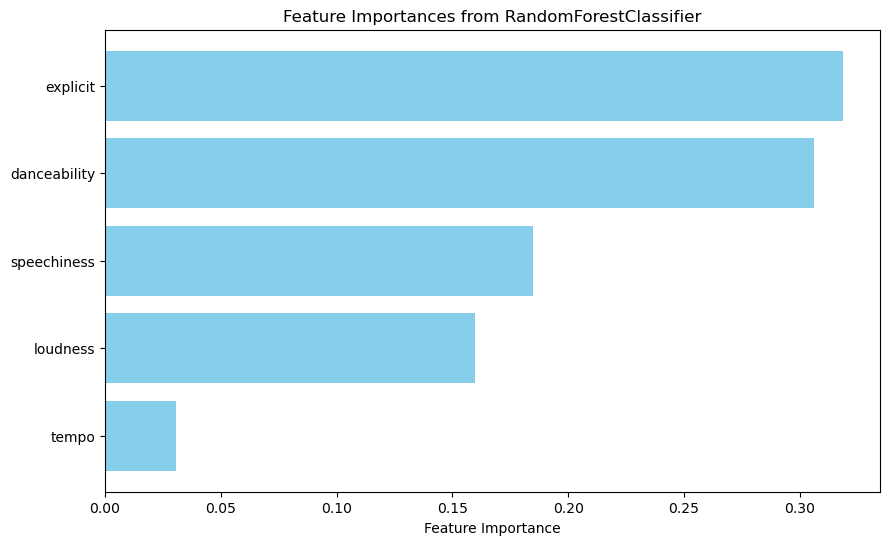


In the following step we investigated Pearson correlation coefficients between the quantitative features we decided to add to the later ml model: the figure shows strong correlations between: energy and loudness (r=0.75) and a strong negative correlation between energy and acousticness (r=-0.72).

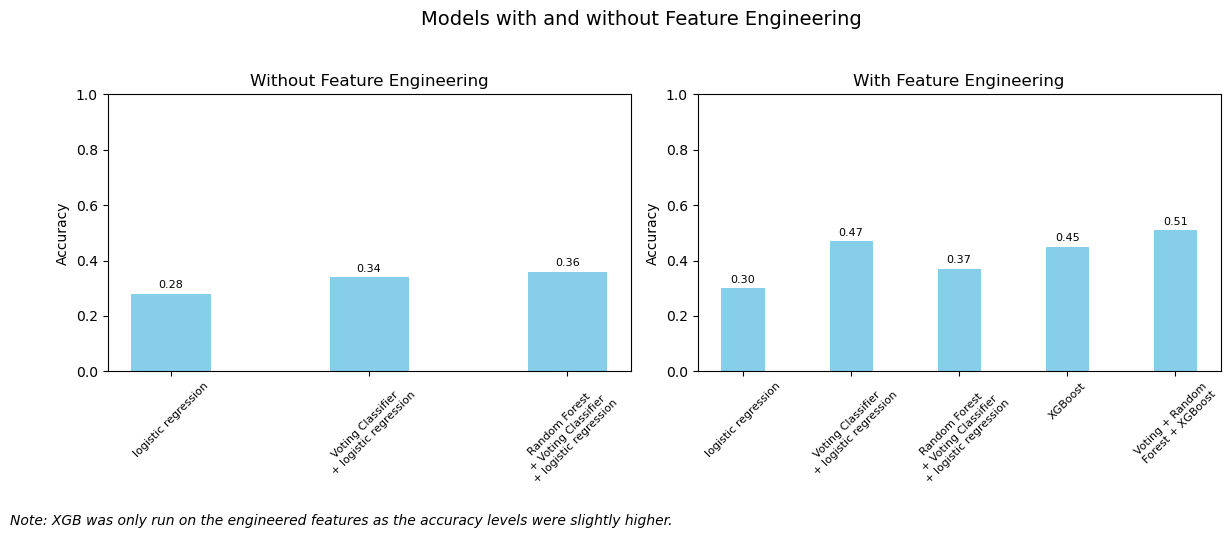


**Models and evaluation:**

In the first step, we split the dataset into training and testing sets (0.3%/0.7%) and used 5 features that exhibited high variance across music genres. We predicted only 5 music genres based on ML models, including Logistic Regression, Voting Classifier, and Random Forest. We assessed feature importance and found that ‘explicit,’ ‘danceability,’ and ‘speechiness’ made the greatest contributions to reducing Gini impurity. The Random Forest model achieved the highest accuracy (0.8) with a Test ROC AUC score of 0.95.

****

Therefore, we decided to run more complex ML models using all music genres (10 music groups comprising all 114 genres as described above), adding more features for a total of 8 features. In each model, we built on the previous model to improve the accuracy score, testing Logistic Regression, Voting Classifier, Random Forest, and XGBoost. We ran each model with and without feature engineering to assess its impact on performance. The graph below shows that feature engineering had a modest effect in boosting model accuracy. However, we ultimately decided to run XGBoost on the engineered features alone. The highest accuracy was achieved when combining all models with XGBoost (0.5), yielding a Test ROC AUC of 0.84.



## **Project Insights and Future Research:**

### 

### The main results of our project show:

### Differences in accuracy scores when using small vs. large datasets for music genre classification.

### A stronger contribution of certain features to reducing Gini impurity in music genre classification.

### Some machine learning models outperform others in terms of accuracy (e.g., Random Forest vs. Logistic Regression).

### The importance of using boosting methods to reduce classification bias.

### The need to address data imbalance for better model performance.

### Modest improvements in model accuracy through feature engineering.

### **Future Research:**

### The strong correlations found between energy, loudness, and acousticness indicate the potential value of performing dimensionality reduction in future studies.

### Exploring additional text-based features or contextual audio data could further enhance model performance.

### 