Analysis of the 200 best Christmas songs (Spotify project)

Capurso Graziana
graziana.capurso@studenti.unipd.it

Di Labbio Daniela
daniela.dilabbio@studenti.unipd.it

Garbin Agata
agata.garbin@studenti.unipd.it

1. Introduction

Spotify's impact on the music industry stems from its use of data-driven insights to enhance user experience. This project focuses on exploring factors influencing track popularity through the analysis of Spotify's extensive dataset. This statistical project uses the Spotify API[1]to analyze the relationship between musical features and audience preferences. Custom features have also been created to provide a more comprehensive framework for understanding how musical features interact with listener preferences.

The statistical project aims to provide insights into the intricate relationship between these features and audience preferences.

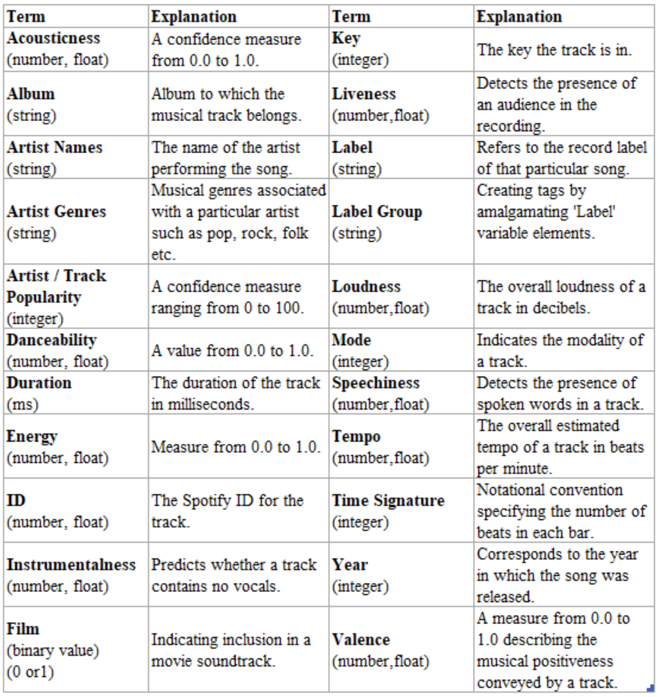
- **Predicting Track Popularity:**
  The investigation will determine if specific feature subsets are related with a track's popularity. Utilizing statistical tools, the project aims to develop a predictive model based on these features.

- **Classifying presence in a film:**
  Beyond these primary objectives, the project extends its focus to classify track quality. This involves developing a classification model to assess a track's suitability for inclusion in a film, considering specific attributes and qualities.

In conclusion through the lens of statistics, we delve into the rich tapestry of Spotify's musical data to unlock valuable insights.

2. Dataset description

Below, we report a list describing the features for each track available in the dataset:



Term	Explanation	Term	Explanation
Acousticness (number, float)	A confidence measure from 0.0 to 1.0.	**Key** (integer)	The key the track is in.
Album (string)	Album to which the musical track belongs.	**Liveness** (number,float)	Detects the presence of an audience in the recording.
Artist Names (string)	The name of the artist performing the song.	**Label** (string)	Refers to the record label of that particular song.
Artist Genres (string)	Musical genres associated with a particular artist such as pop, rock, folk etc.	**Label Group** (string)	Creating tags by amalgamating 'Label' variable elements.
Artist / Track Popularity (integer)	A confidence measure ranging from 0 to 100.	**Loudness** (number,float)	The overall loudness of a track in decibels.
Danceability (number, float)	A value from 0.0 to 1.0.	**Mode** (integer)	Indicates the modality of a track.
Duration (ms)	The duration of the track in milliseconds.	**Speechiness** (number,float)	Detects the presence of spoken words in a track.
Energy (number, float)	Measure from 0.0 to 1.0.	**Tempo** (number,float)	The overall estimated tempo of a track in beats per minute.
ID (number, float)	The Spotify ID for the track.	**Time Signature** (integer)	Notational convention specifying the number of beats in each bar.
Instrumentalness (number, float)	Predicts whether a track contains no vocals.	**Year** (integer)	Corresponds to the year in which the song was released.
Film (binary value) (0 or1)	Indicating inclusion in a movie soundtrack.	**Valence** (number,float)	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.

Figure 1. Features abbreviation meaning

[1] https://developer.spotify.com/documentation/web-api/reference/get-audio-features

3. Explorative Analysis

We prepared the dataset and then examined it in its main aspects in further depth.

3.1 Distribution of Variables

As we explore the distribution of musical features scored from 0 to 1, we uncover some intriguing patterns that shed light on the musical landscape of Christmas songs, as shown in Figure 2.

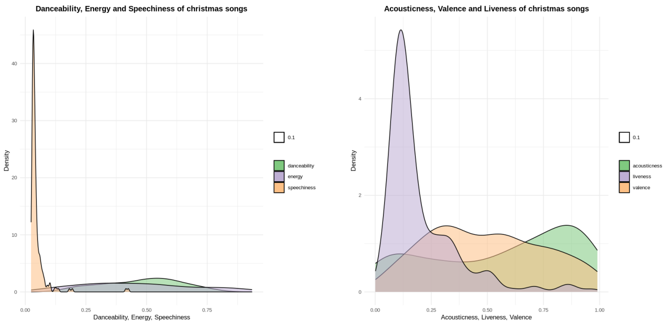


Figure 2: Density plot

First and foremost, we notice that a majority of the songs tend to have a *danceability* range between 0.45 and 0.75, which means that they are neither too slow nor too fast. Interestingly, the distribution of *energy* across the songs appears to be quite uniform, indicating that Christmas music tends to maintain a consistent level of energy throughout. Another key observation is the prevalence of tracks with *speechiness* between 0 and 0.1, which suggests that many Christmas songs have fewer lyrics. This finding is not surprising, given that instrumental music and melodies are often a hallmark of Christmas songs. Moving on to *acousticness*, we observe a right-skewed distribution, which means that highly acoustic songs are more prevalent in the Christmas music genre. These songs typically feature orchestral instruments, an unaltered voice, acoustic guitars, and natural drum kits, which help create a warm and authentic sound. In contrast, less acoustic Christmas songs tend to incorporate synthesizers, electric guitars, and amplified instruments, which give them a more modern and upbeat feel. Lastly, we note that the *valence* variable has a peak of density between 0.25 and 0.70, indicating that most Christmas songs are not overwhelmingly positive. While Christmas is often associated with joy and happiness, many Christmas songs have a more reflective and nostalgic tone that is not necessarily upbeat.

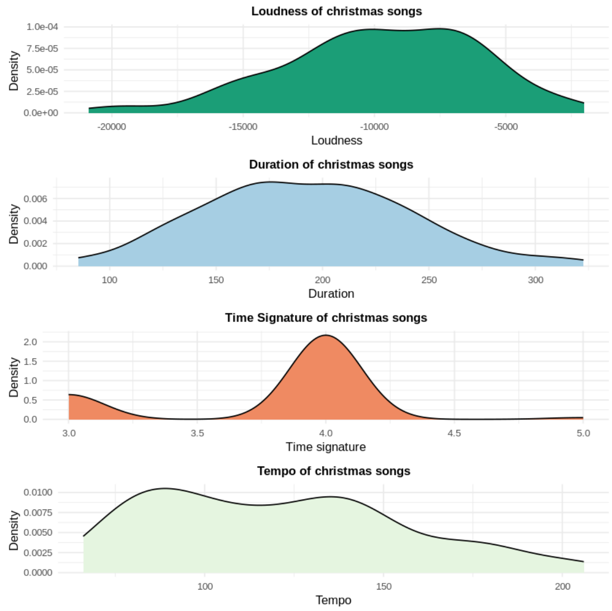


Figure 3: Density plot

As shown in Figure 3, the majority of tracks have a moderate volume. In terms of *duration*, they range evenly from 100 to 270 seconds, offering diverse lengths for different moods. Many tracks use a familiar four-beat *time signature*, ensuring a steady rhythm. When it comes to *tempo*, most fall slightly below 100 or just under 150 beats per minute (BPM). In summary, the volume, length in seconds, time signature, and tempo distribution of the tracks offer a well-rounded listening experience that caters to a wide range of preferences.

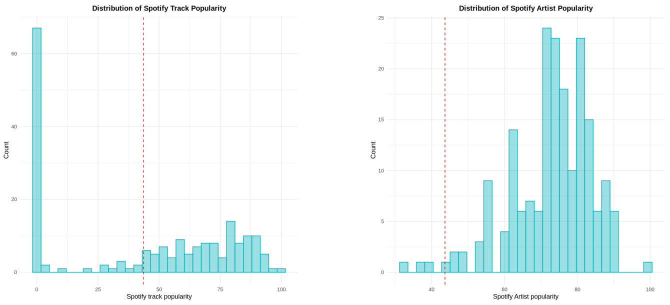


Figure 4: Tracks and artists popularity

To complete our exploration, first of all, we look at the distribution of **track popularity** and we see that there's an average popularity range of 40-50, with numerous tracks having a lower popularity around 0. Regarding **artist popularity**, the average lies within 40-50, with a substantial number of tracks exhibiting a popularity around 70, as determined by Spotify.

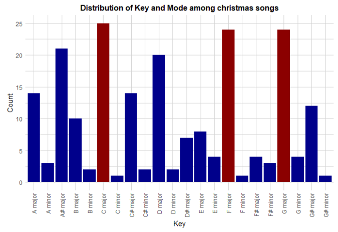


Figure 5: Key Signature

Finally based on this plot we derive that the most popular keys among tracks are G major, C major and F major. Since major keys are perceived as happy and minor keys as sad, it is logical that they dominate the distribution according with the very wide peak of valence along the tracks.

3.2 Data organization

Additionally, we check for correlations between variables and plot them. The variables **loudness** and **energy** are highly correlated, so we decide to remove the second one. We have also plotted the correlation between genres in order to merge the similar ones or the ones that belong to just few songs and make the analysis more significant.

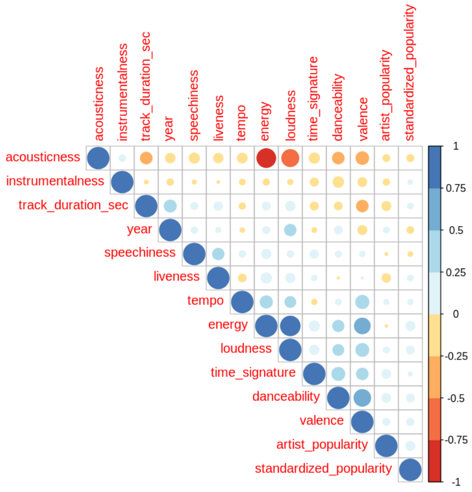


Figure 6: Correlation of numeric variables

4. Model Development

In this section, the methodology used to solve each of the tasks will be discussed. For all the tasks discussed below, the train test split employs a split ratio of 0.8.

4.1 Linear Model

The initial predictive model for song success utilizes linear regression, a highly effective method for forecasting a numerical dependent variable based on independent variables. The model undergoes a rigorous selection process, including backward, forward, and hybrid methods, resulting in a highly accurate model with varied influencing factors on the popularity of a track. The forward regression method revealed a model with key predictors, emphasizing the paramount importance of features like, **soul**, **easy listening**, **artist popularity**, **year** and **liveness** in determining track popularity. The model's adjusted R-squared of 0.8703 suggests a moderate fit, but it still captures approximately 87% of the variability in track popularity, making it an invaluable tool for predicting song success. The backward regression technique unveiled a model that highlights the crucial significance of certain predictors, underscoring the importance of features such as

artist popularity, **acousticness**, **instrumentalness**, **liveness**, **year**, **label group**, **easy listening**, **motown** and **soul**. We can see that in comparison to the previous subsection selection procedure, the adjusted R-squared of the model is lower. Specifically, the R-squared value is 0.8154. The stepwise hybrid approach did not introduce additional insights beyond backward selections, emphasizing the consistency in variable importance. This consistency validates the selected variables' significance, showcasing the importance of certain audio features in predicting track popularity. We then delved deeper into the variance explained by each principal component and decided to retain the first 12 principal components. Next, we utilized these selected principal components to construct a linear regression model that incorporated the response variable track popularity from the training data. Our results showed that the model had an adjusted R-squared of 0.2267, indicating that it could explain 22.67% of the variability in the response variable.

4.2 LOESS

Local regression, employs local fitting to create a regression model for a given dataset. We employed cross-validation to determine the optimal `span` value for this method. Consequently, for our ultimate LOESS regression model, we have selected a `span` value of 0.1, employed as an argument within the `loess()` function. As predictors, we've chosen the most significant variables derived from the stepwise linear regression: **artist popularity** and **soul**. To evaluate the model's performance, we used these predictors to make predictions on the test set, enabling the calculation of error measures.

4.3 Leaps

To kick off our analysis, we started by building a linear model to predict the target variable track popularity using all available variables. However, we soon discovered a problem with "linear dependencies", which could skew our results. To address this, we calculated variance inflation factors (VIFs) to identify any variables with multicollinearity issues. Once we had identified the problematic variables, we removed them from our analysis. This allowed us to proceed with the `leaps` package, which can handle models with fewer than 31 variables and avoids linear dependencies. To select the best variables for our model, we used the `leaps()` function and the Mallow's (Cp) criterion. We ran several iterations to strike a balance between model complexity and fit, observing how the value of Cp changed as we adjusted the model. The variables that had been selected are **loudness**, **instrumentalness**, **track duration**, **film**, **easy listening**, **mellow gold**, **motown** and **altro** (this variable referres to those subgenres that do not fall into the other classes). To evaluate our models, we used leave-one-out cross-validation (LOOCV) to calculate prediction errors for each one. We chose the model with the minimum cross-validation error, which became our final model. We then applied this model to the test data, calculating mean square errors (MSE) and mean absolute errors (MAE) to assess its predictive performance. The model that was developed to predict the target variable was thoroughly assessed and validated using LOOCV. The results showed that it had excellent predictive ability on the test data, as evidenced by the low MSE and MAE values.

4.4 Ridge and Lasso

In this task, after properly joining the dataset and data cleaning, different regression models such as Ridge and Lasso are used. All the models are fitted with `glmnet` package and for the selection of the hyper-parameter, the `cv.glmnet` is used. Finally, the models are fitted with the best hyper-parameters selected by the cross-validation and then prediction is made based on the newly fitted model. As for the assessment of the model, the mean square error and the absolute mean square error are used. The variables identified and chosen by Lasso include **artist popularity**, **key signature**, **liveness**, **label group**, **easy lis-**

tening, **motown**, **pop** and **soul**, meaning that they hold significant relationships with *track popularity*.

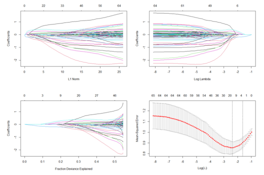


Figure 7: Lasso

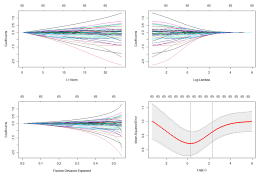


Figure 8: Ridge

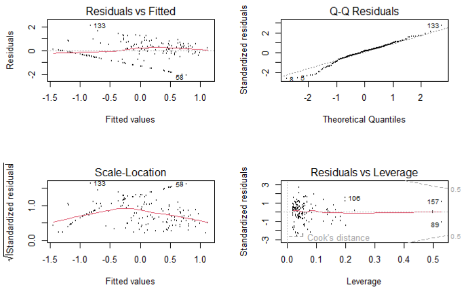


Figure 9: Best GLM

4.5 LARS

Least-angle regression (LARS) is an algorithm for fitting linear regression models to high-dimensional data. Suppose we expect a response variable to be determined by a linear combination of a subset of potential covariates. Then the LARS algorithm provides a means of producing an estimate of which variables to include, as well as their coefficients. In our case, subsequent to fitting the model on the training set and generating predictions on the test set, we employed cross-validation to determine the precise number of variables utilized by the model that yields the lowest mean squared error (MSE). The consistent findings indicate that the optimal number of variables to use is 3 (**soul**, **label group** and **pop**), achieving an RSS of 144.87 (not so satisfactory) and a Cp of 4.

4.6 Generalized Linear Models

After apply lasso regression, which helped us select pertinent variables while tackling any multicollinearity and linear dependence problems, we turned to the `bestglm` package to identify the ideal model for our needs based on Akaike's information criterion (AIC). Thanks to this approach, we were able to pinpoint the most relevant variables, including **artist popularity**, **key signature**, **liveness**, and **label group**, among many others.

4.7 Support Vector Machine

We concluded our task by framing a classification problem using the SVM (Support Vector Machine) algorithm. Initially, we applied the `gcdnet` function to the model matrix. Subsequently, employing cross-validation helped us pinpoint the minimum lambda value, which corresponded to the significant variables necessary for our model. Consequently, the final set of variables utilized for predictions on the test set comprises: **danceability**, **loudness**, **acousticness**, **liveness**, **tempo**, **duration**, **year**, **swing**, **mellow gold**, **pop**, **folk**, and **altro** variables. The accuracy achieved stood at 0.70.

5. Results

5.1 Track - Popularity prediction

From the prediction on track popularity we can conclude that a popular Christmas song is principally from a very popular artist and it is not classificable as a soul song. When it comes to selecting a method for regression analysis, it's crucial to evaluate each option's performance. In this case, we've assessed seven different methods based on their Mean Squared Error (MSE) and Mean Absolute Error (MAE) values.
As we can see in the Table 1 Backward and Hybrid stood out with similar MSE and MAE values, indicating that they performed similarly well. They both outperformed the Forward method, making them promising options. On the other

Model	MSE	MAE
forward	0.7207534	0.6779238
backward	0.5278944	0.5361365
both	0.5278944	0.5361365
pc	2432.6959289	45.9212121
loess	0.6380030	0.6524411
leaps	0.8798283	0.7990952
cv	0.8735904	0.7958484
bestglm	0.5674993	0.5690745
lasso	0.6825470	0.6987524
ridge	0.6114169	0.6572209
lars	0.9507087	0.8685130

Table 1: Mean Squared Error (MSE) and Mean Absolute Error (MAE)

hand, Principal Components (PC) had significantly higher MSE and MAE values, which may not make it the best choice for our specific task. However, it's important to note that each method's suitability depends on the data's characteristics and requirements. Moving on to Leaps and Cross-Validation (CV), these methods had relatively higher MSE and MAE values than Backward and Hybrid, but they could still be reasonable choices. Bestglm, Lasso, and Ridge showed relatively lower MSE and MAE values, indicating better performance than some other methods. Lastly, Lars had the highest MSE and MAE values among all, which suggests potentially poorer performance in our particular context.

5.2 Film Classification

In classification SVM model analyzing the presence of a Christmas song in a film, certain features demonstrate a significant influence. Specifically, acousticness and liveness, along with the swing genre, exhibit a positive correlation with the presence of a Christmas song. Conversely, features such as danceability, loudness, tempo, duration, year, and specific genres like pop, mellow gold, and 'altro' showcase negative coefficients. This indicates that these attributes tend to have a diminishing effect on the likelihood of a Christmas song being present in a film within the context of this SVM classification model. To evaluate its performance we have printed the following confusion matrix:

	No	**Yes**
-1	9	5
1	45	106

Table 2: Confusion Matrix

In addition, we have calculated other metrics such as precision, recall, and F1 score, which are reported below.

Metric	Value
Accuracy	0.6969697
Precision	0.954955
Recall	0.7019868
Score F1	0.8091603

Table 3: Metrics Results