**PROJECT REPORT**

Aashreya Reddy Pasham(axp190018)

**Executive Summary**

Predicted mainstream hit music based on song features using an ensemble learning method, Random Forest Model with 10 fold cross validation to achieve an accuracy of 86%.Performed k-mean clustering to find different types of music such as acoustic, metal, mellow and mainstream music. Predicted the clusters using a multinomial logit model to attain an accuracy of 99.5%. Built an interactive shiny application for providing song recommendations based on the user’s input and also created an interactive EDA and model building. Mainstream hit music not only depends on the features of music but also based on the Artist. All songs of popular artists would be on the billboard list at least once but that doesn’t necessarily make it a hit song. It can depend on how long it stays on the billboard list which is something that can be looked into for future analysis.

1. **Introduction**

The Billboard Hot 100 is the music industry standard record chart in the United States for songs, published weekly by Billboard magazine. Chart rankings are based on sales, radio play, and online streaming in the United States. These songs are considered to be a mainstream hit in the dataset which contain features of all the songs from Jan 1st,2010, ending at Dec 31st, 2019 which is extracted using a Spotify web API.

1. **Data Description**

This dataset contains 6398 observations and 19 variables described below.

* track: The Name of the track.
* artist: The Name of the Artist.
* uri: The resource identifier for the track.
* danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
* energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
* key: The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C?/D?, 2 = D, and so on. If no key was detected, the value is -1.
* loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality

of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.

* mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
* speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
* acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. The distribution of values for this feature look like this:
* instrumentalness: Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. The distribution of values for this feature look like this:
* liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
* valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
* tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
* duration\_ms: The duration of the track in milliseconds.
* time\_signature: An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
* chorus\_hit: This the the author's best estimate of when the chorus would start for the track. Its the timestamp of the start of the third section of the track. This feature was extracted from the data received by the API call for Audio Analysis of that particular track.
* sections: The number of sections the particular track has. This feature was extracted from the data received by the API call for Audio Analysis of that particular track.
* target: The target variable for the track. It can be either '0' or '1'. '1' implies that this song has featured in the weekly list (Issued by Billboards) of Hot-100 tracks in that decade at least once and is therefore a 'hit'. '0' Implies that the track is a 'flop'.

1. **Preprocessing Data**

The dataset is balanced; contains 3199 1’s and 3199 0’s in the Target variable. It’s a clean dataset with no missing values or incorrect inputs.

1. **Exploratory Data Analysis**

Unique number of artists in the dataset are 3355. Different kind of keys used are 1, 5, 9, 0, 2, 7, 8, 11, 3, 4, 6, 10 and time\_signature are either 0, 1, 3, 4 or 5.

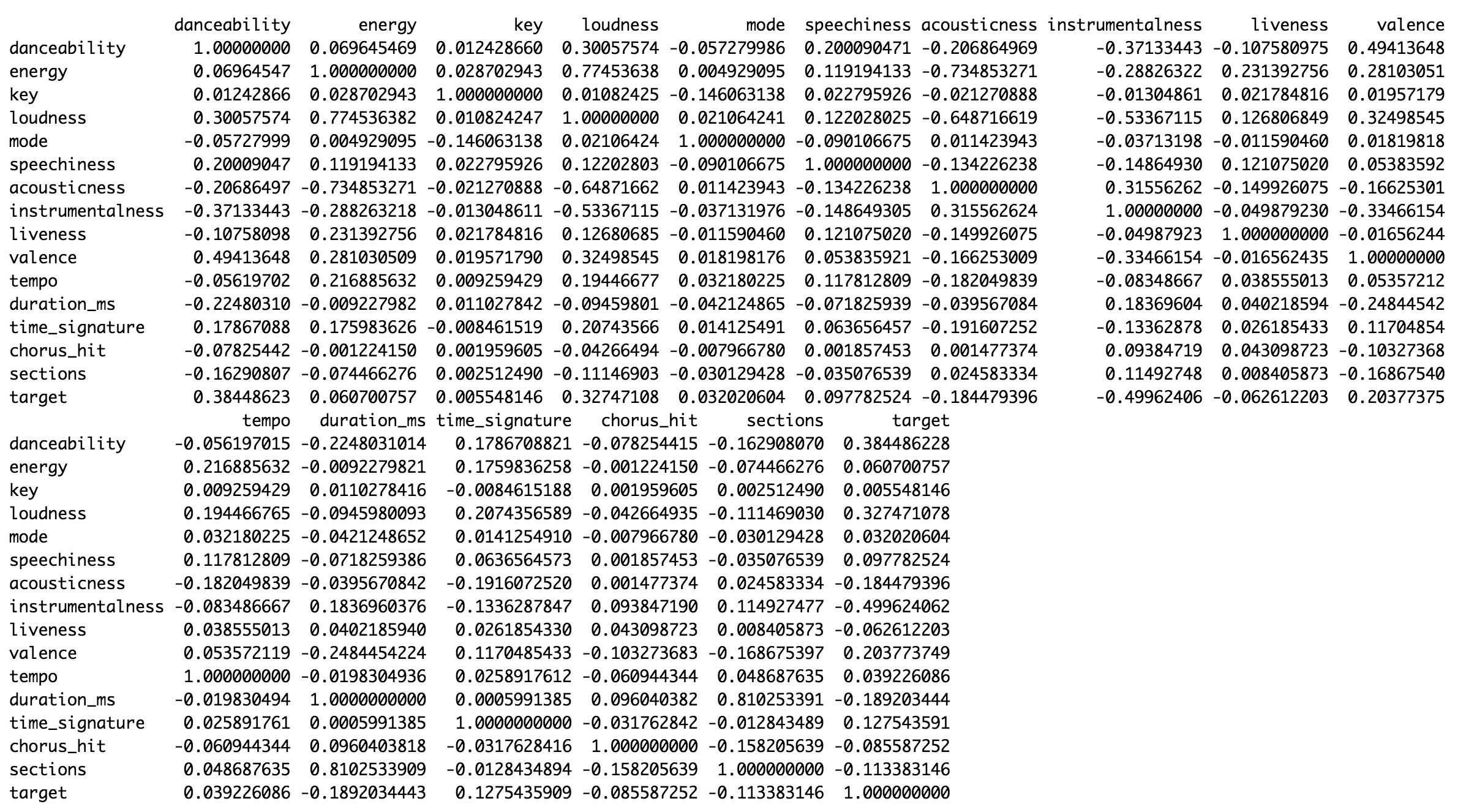


Fig 1.1 – Correlation Table

Looking at the fig 1.1 we can tell that danceability, loudness, acousticness, instrumentalness, valence, time\_signature variables are correlated with the target variable so the could be significant in predicting it. Can also see how much the features are correlated with each other for instance, danceability has a 0.30 correlation with loudness. we can also see which features can be eliminated as well. Key and mode have the least correlation with the target variable.

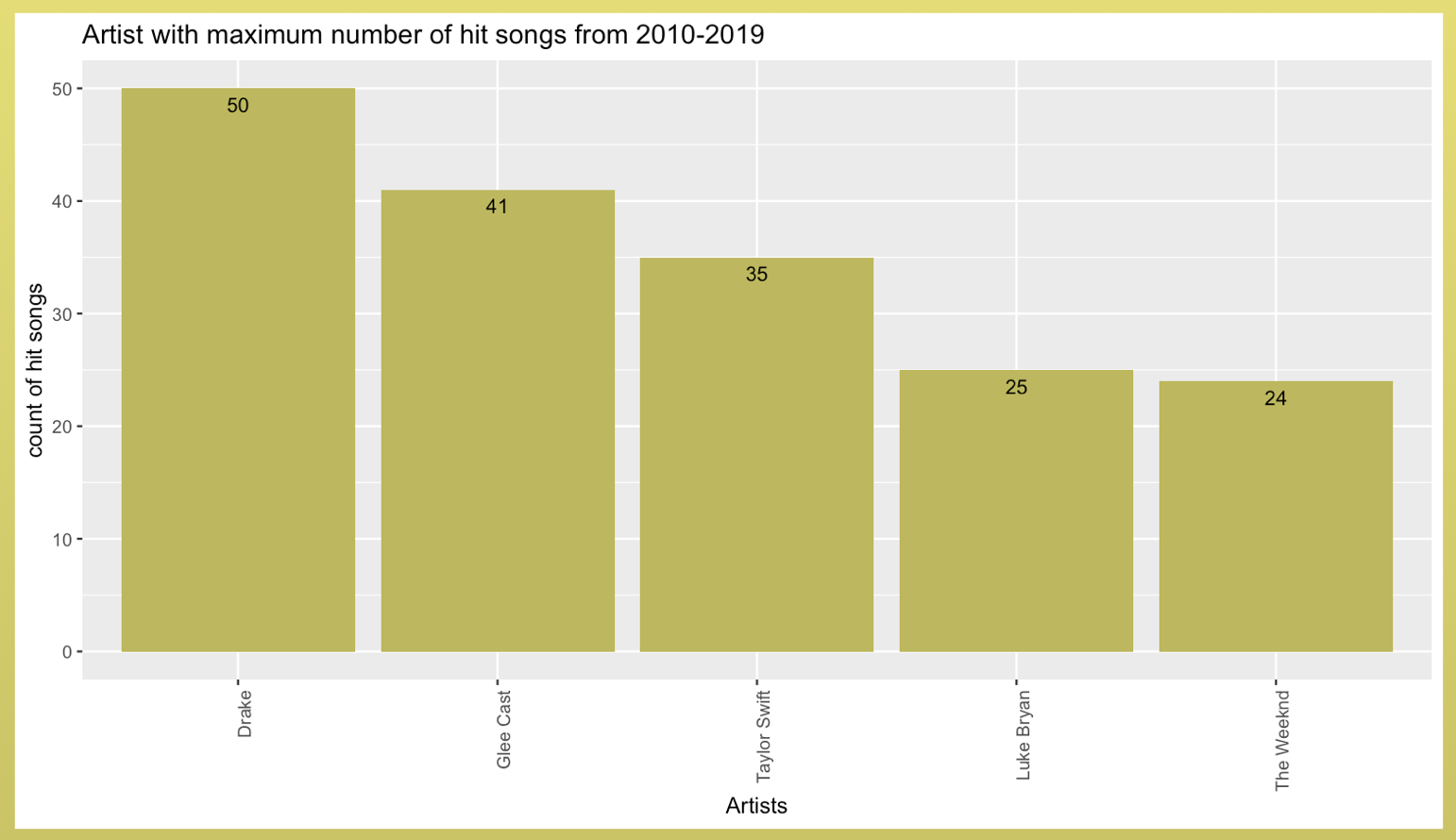


Fig 1.2- Artist with maximum hit songs(excluding featured songs)

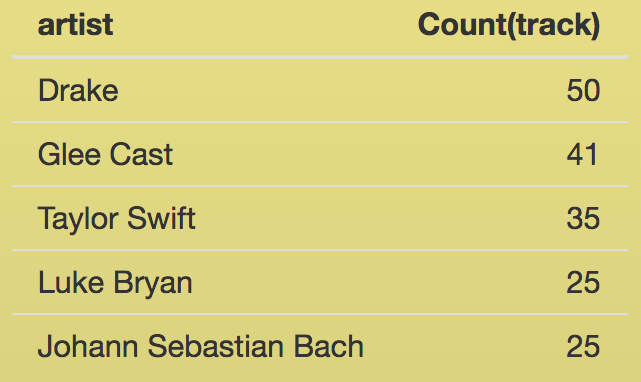


Fig 1.3- Songs per Artist( Excluding featured songs).

Looking at fig 1.2 and 1.3 we can notice that all of Drake’s songs are hits which is unlikely but probably was on the billboard list for a short period when the album was dropped and since it’s a very popular

artist. Mainstream hits don’t depend just on the song features but also based on the popularity of the Artists.

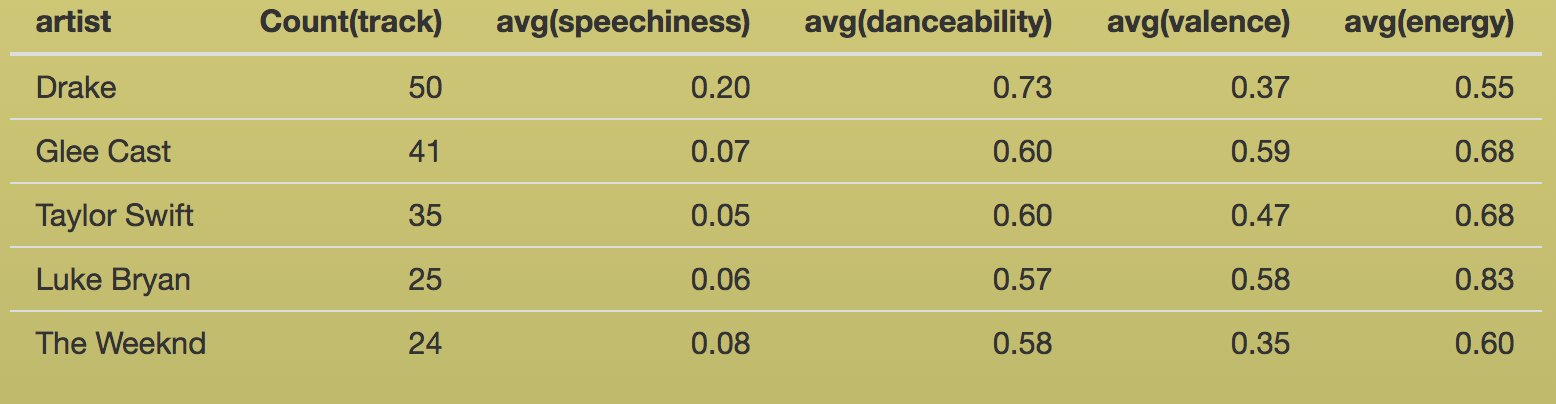


Fig 1.4- Average features of the top Artists of the Decade.

Looking at Fig 1.4 we can notice that there is a clear difference in speechiness in hip hop artists such as Drake compared to pop artists such as taylor swift.

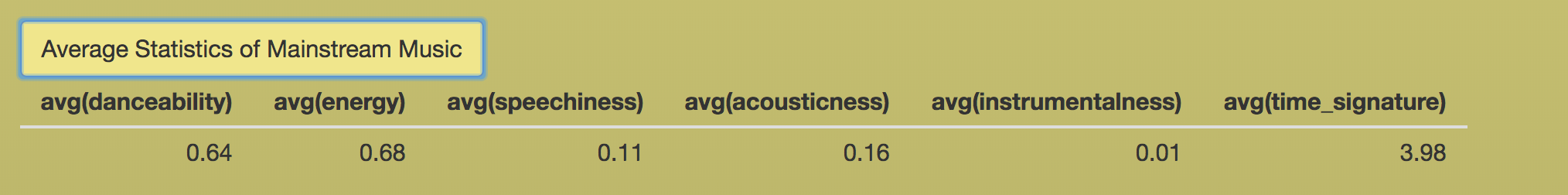


Fig 1.5: Average features of all mainstream hit music.

Looking at Fig 1.5 you notice that all the hit music on an average has high danceability, energy, speechiness to be medium and to be more acoustic than instrumentalness.

1. **Empirical Analysis**

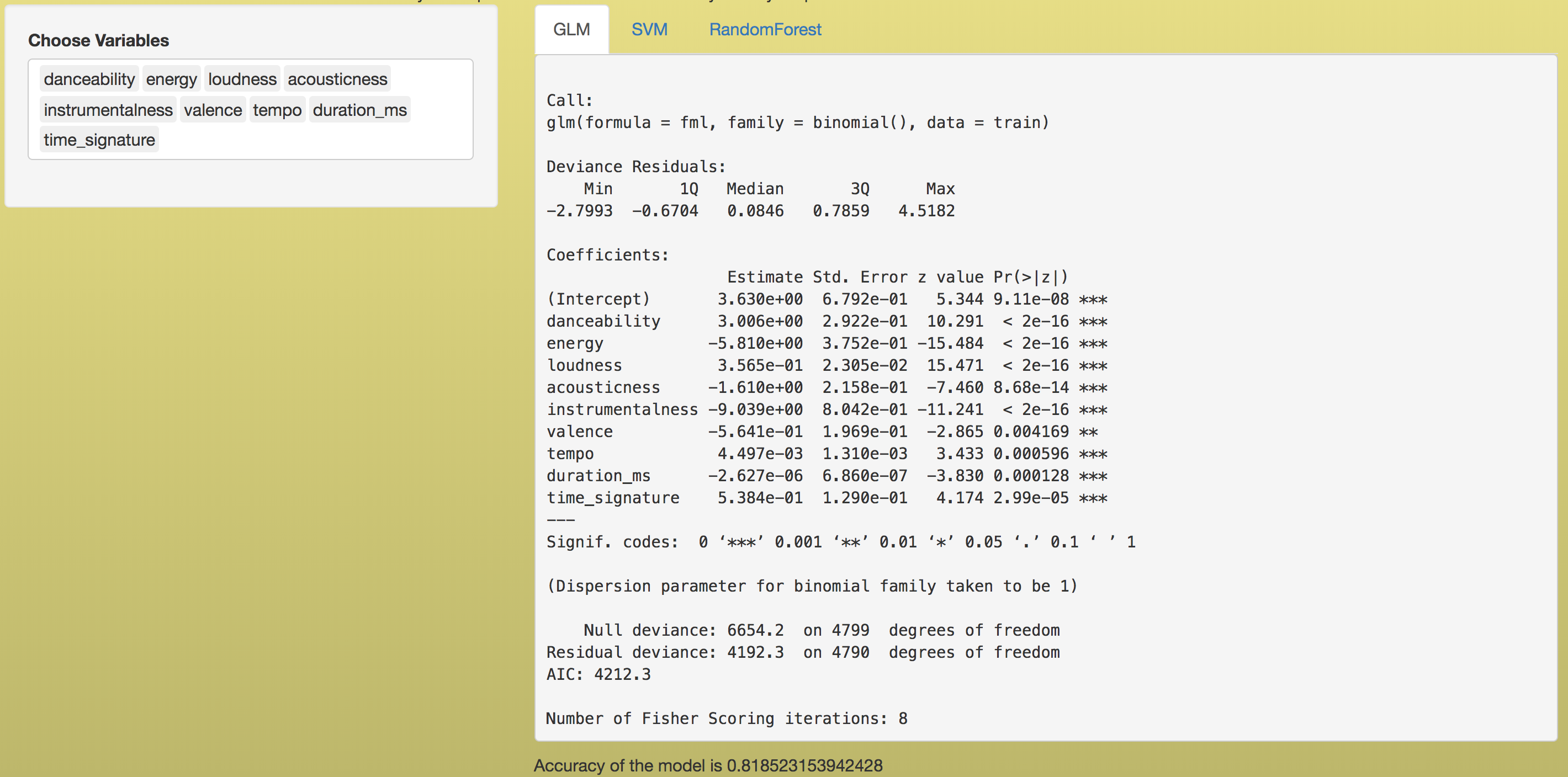
****

Fig 2.1- Interactive GLM model with summary and accuracy

****

Fig 2.2- Interactive SVM model with summary and accuracy

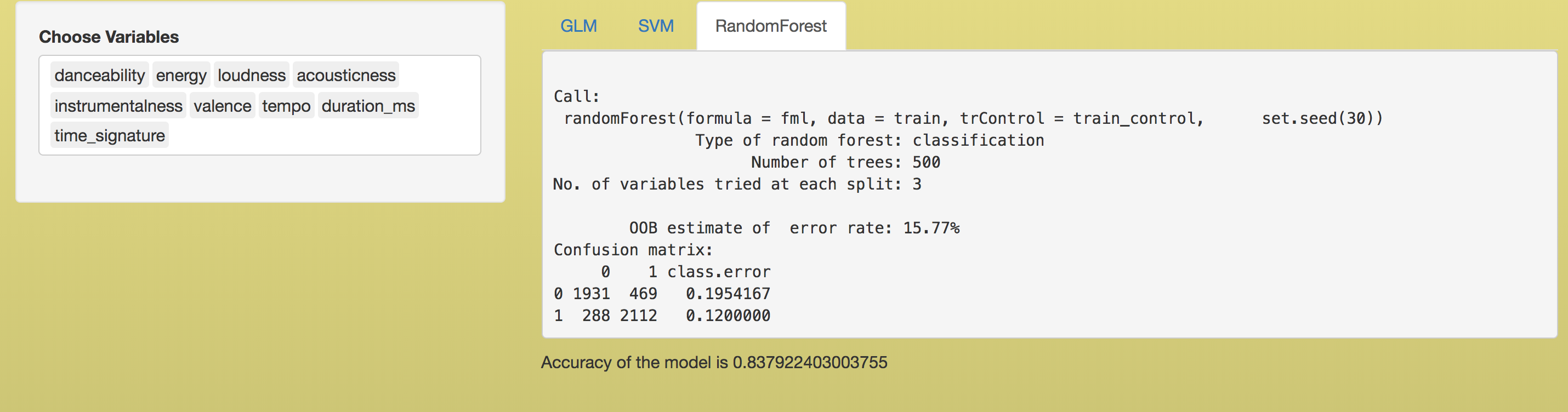


Fig 2.3- Interactive random forest model with summary and accuracy

The goal here is to predict if a track is going to be a mainstream hit or flop given it's features. Looking at the Fig 2.1 we can tell which variables are significant in predicting the target variable. Danceability, acousticness, loudness, valence, tempo, duration, time\_signature, energy and instrumentalness are significant variables in predicting the target variable. I've build svm and randomforest model using these significant variables as the independent variables and achieved a higher accuracy compared to GLM. The highest accuracy achieved is around 85% using Random forest algorithm with a 10-fold cross validation.

Made all the models interactive where you can select the variables used to predict the target variable and look over the model summaries and its accuracies.

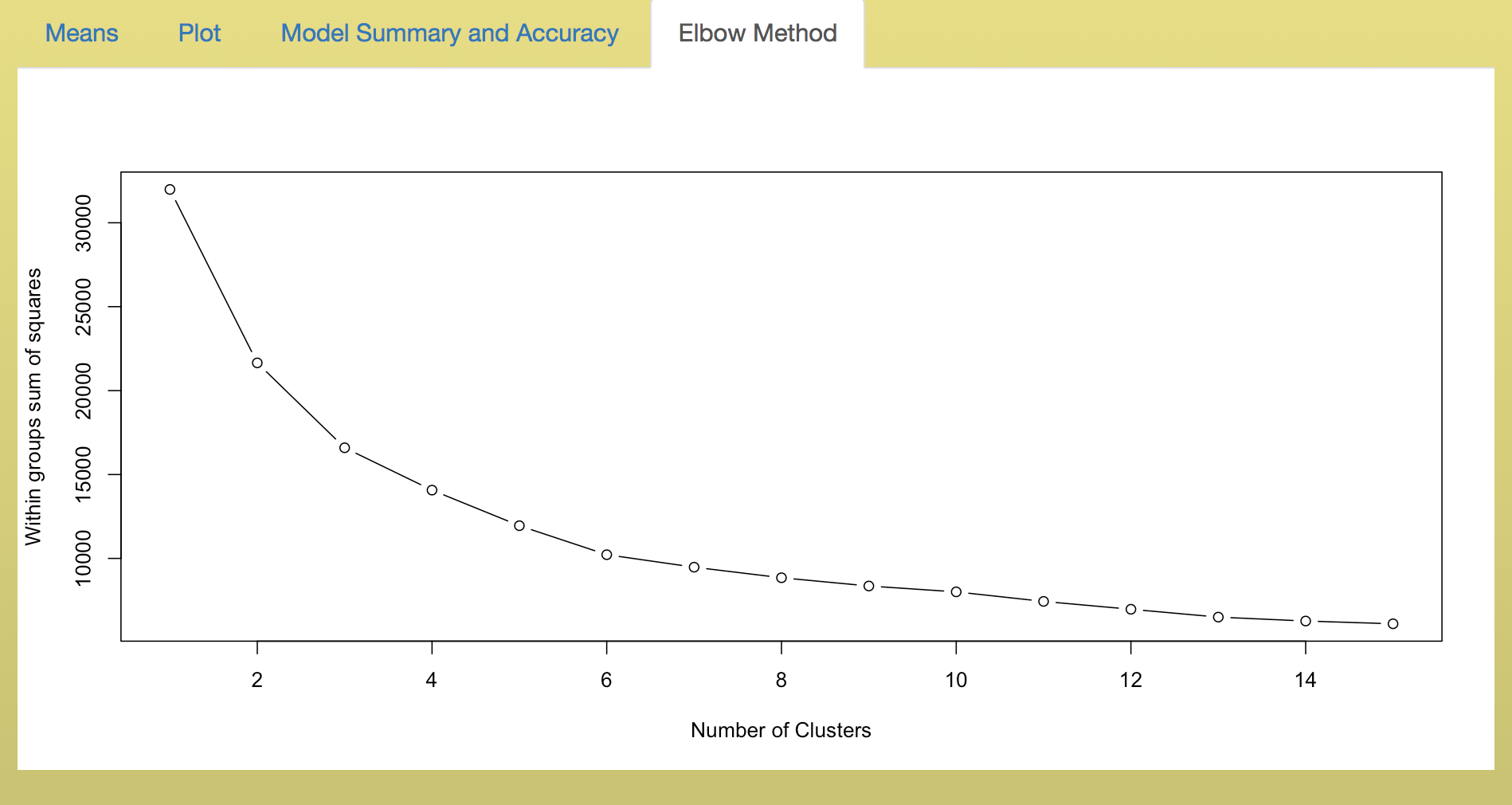


Fig 2.4 – Elbow Method

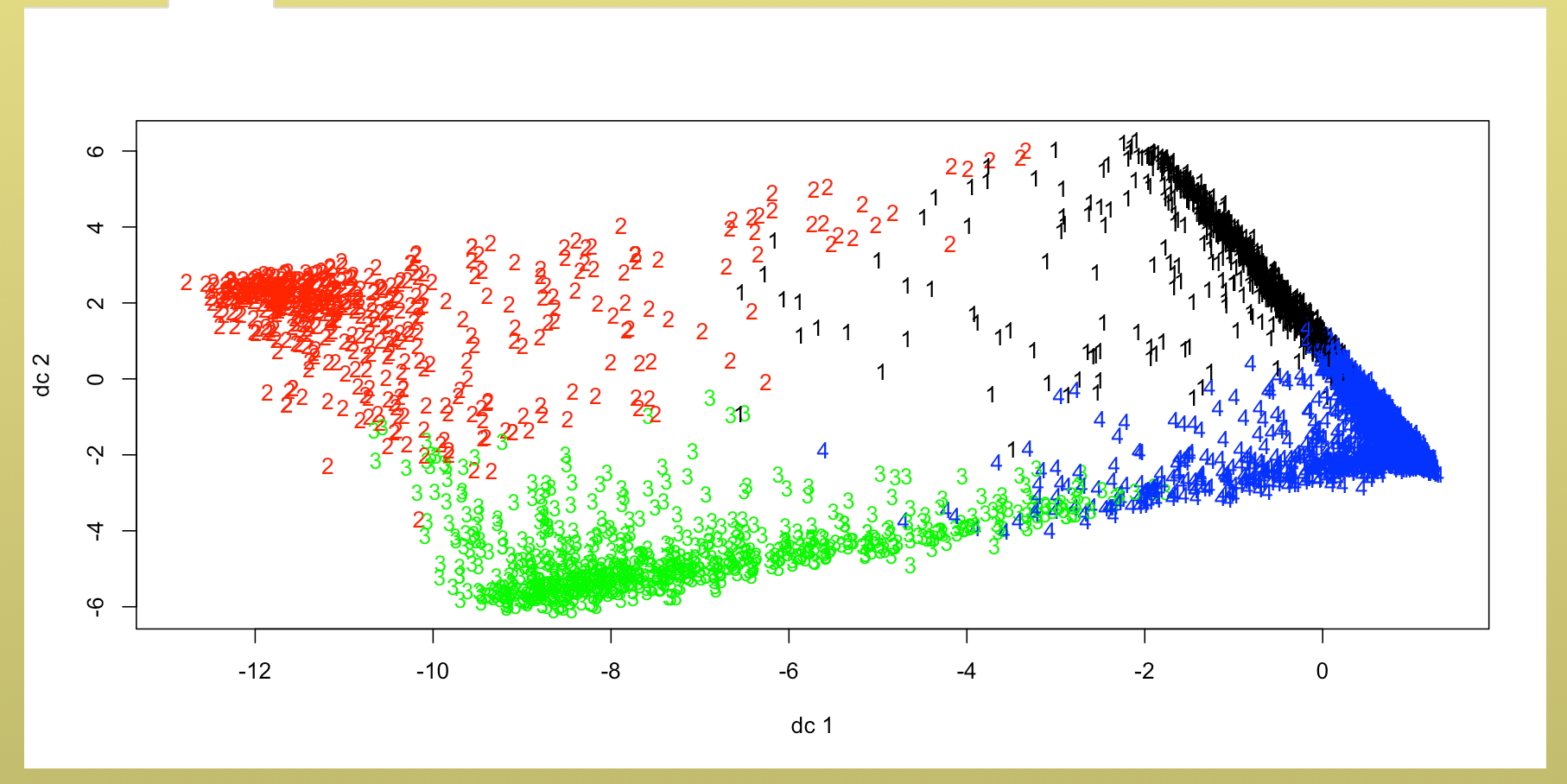


Fig 2.5- Clusters plot

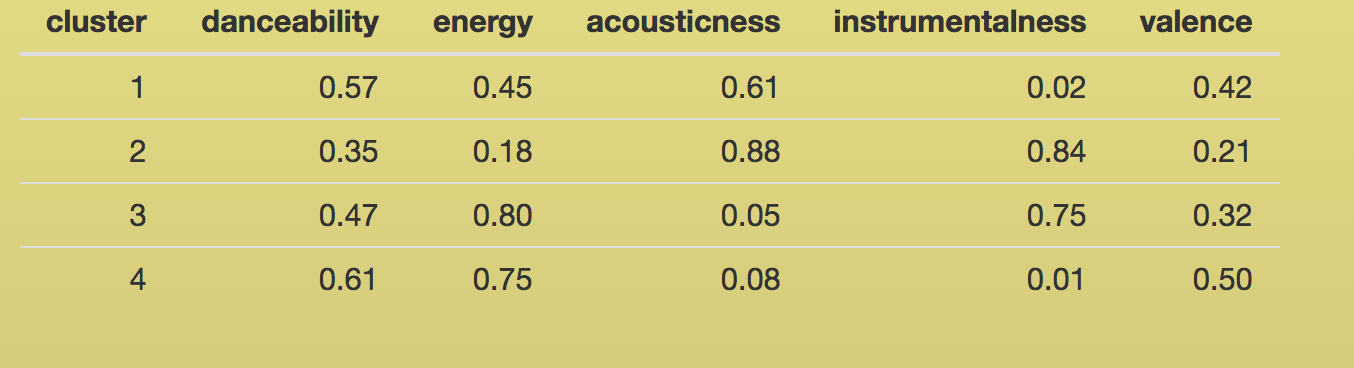
****

Fig 2.6- Mean values of all the clusters

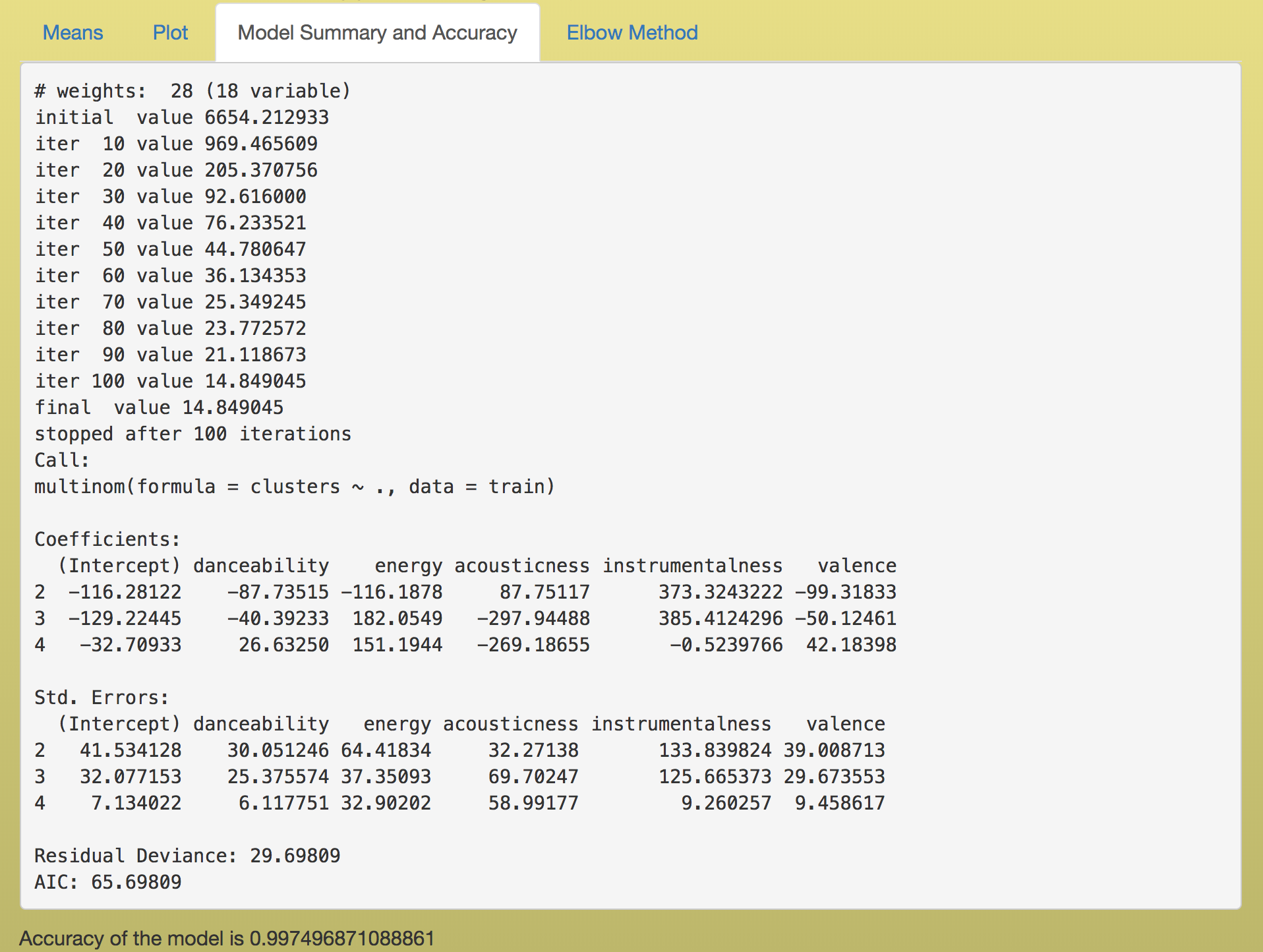


Fig 2.7- Multinomial model summary and Accuracy

Fig 2.4 elbow method and Fig- 2.5 helps identify the optimal number of clusters required. The variables I've selected our danceability, energy, acousticness, instrumentalness and valence. The ideal number of clusters are 4.

Fig 2.6 tells us that Cluster 1 has medium danceability, energy, valence but high acounsticness and low instrumentalness. (Acoustic Music). Cluster 2 has low danceability, energy, valence but very high accousticness and instumentalness (mellow music).Cluster 3 has medium danceability,valence but high energy, instrumentalness and low accousticness (metal music).Cluster 4 has high danceability, energy, medium valence but low acousticness and instrumentalness like mainstream music so could have a lot of hits based on previous analysis. Achieved an accuracy of 99.7%.

Fig 2.7 Gives us the model summary and its accuracy in predicting which cluster a track belongs to based on it’s features (danceability, energy, acousticness, instrumentalness and valence).

You get to select the number of clusters and the variables for clustering and all the tabs in Fig- 2.4, 2.5, 2.6, 2.7 change accordingly.

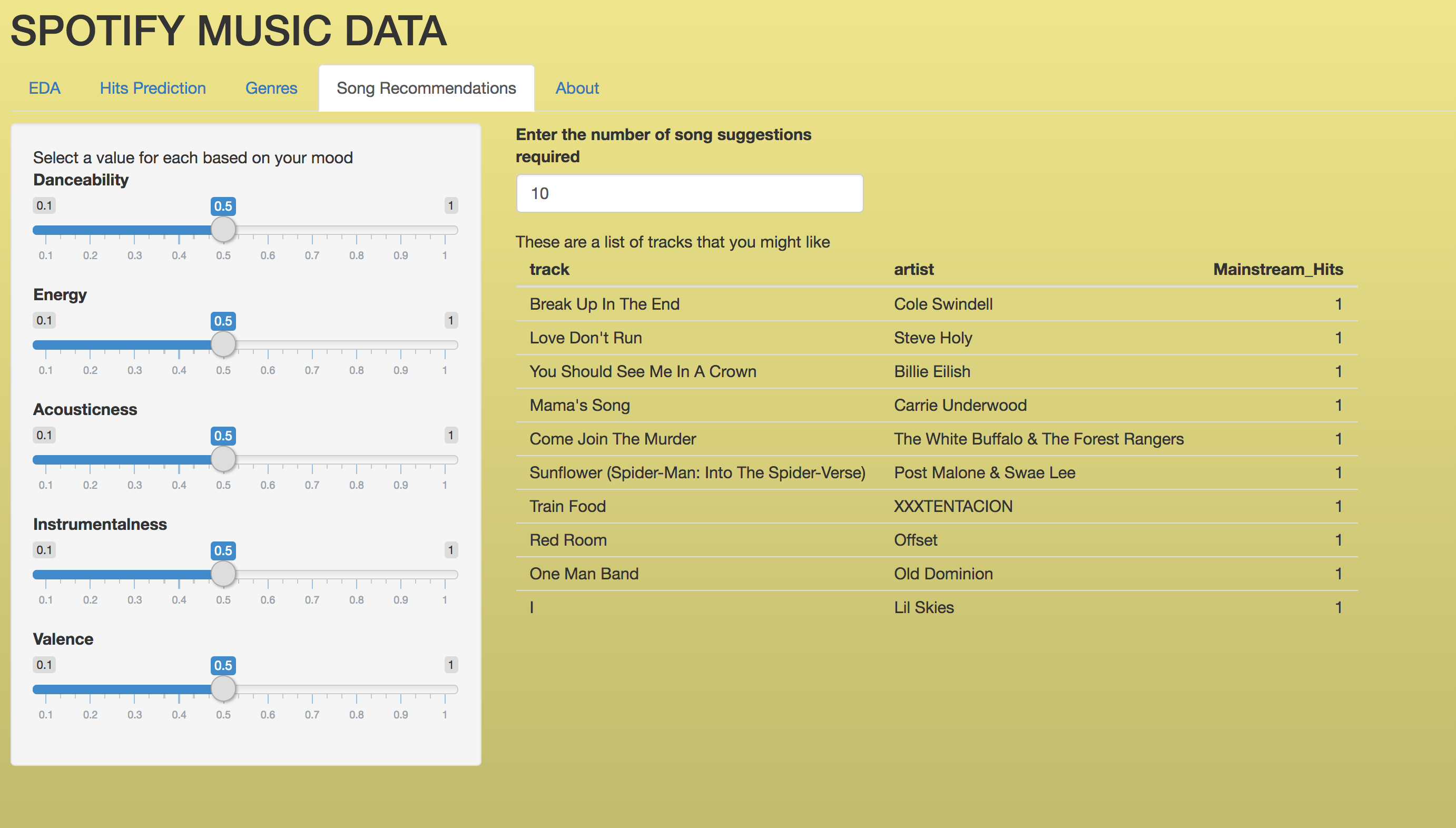


Fig 2.8- Song Recommendations

Fig 2.8 is also an interactive shiny application to give song recommendations based on the user choice of music features. Considering all the important features(danceability, energy, acousticness, instrumentalness and valence) and using 6 clusters, we predict which cluster do these user inputted features fall under(Accuracy 99.3%) and provide a list of songs that belong to that cluster starting with all the mainstream hits of that cluster following by the unpopular songs. You can enter the number of song suggestions required ranging from 0 to 50.

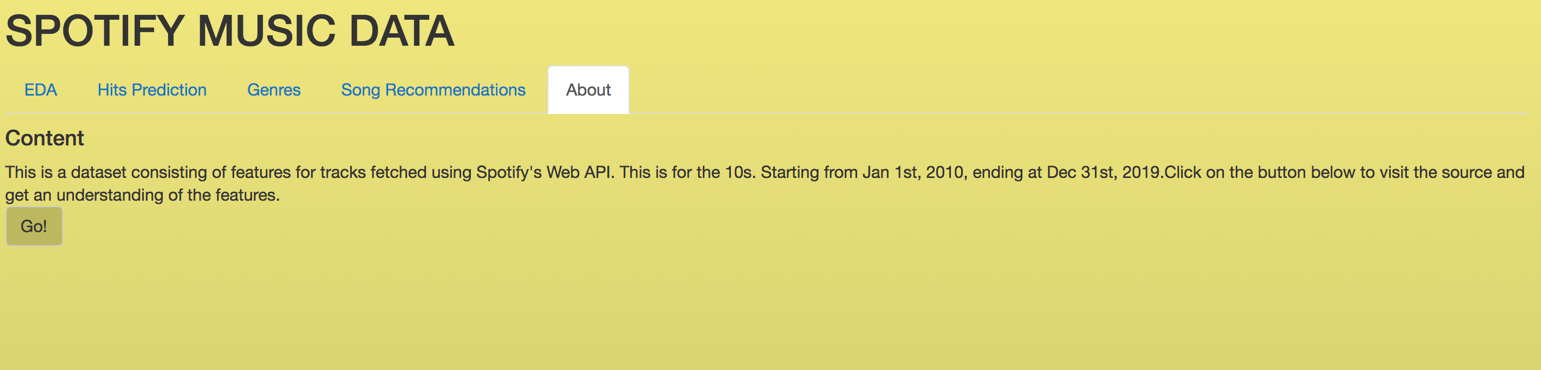
****

Fig 2.9- About

Fig 2.9 is a tab that provided you information about the dataset and when you click on the button “GO!” it redirects you the source which is <https://www.kaggle.com/theoverman/the-spotify-hit-predictor-dataset>. It gives you all the information required to understand the dataset.

1. **Conclusions**

Using a large dataset , predicted if a song is going to be a hit or a flop based on its music features. Clustered to find different genres of music and provided song recommendations based on user’s Input and also concluded that a mainstream hit does not just depend on the features but also be based on the popularity of the artist.

1. **Sources**

<https://shiny.rstudio.com/>

<https://shiny.rstudio.com/images/shiny-cheatsheet.pdf>

<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans>

<https://www.rdocumentation.org/packages/e1071/versions/1.7-3/topics/svm>

<https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest>

<https://www.w3schools.com/sql/>

<https://stats.stackexchange.com/questions/61090/how-to-split-a-data-set-to-do-10-fold-cross-validation>

<https://www.r-bloggers.com/finding-optimal-number-of-clusters/>