# Inferring skip patterns from Spotify's soundtrack recommendation

*Ravitashaw Bathla*

---

## Summary

A study of the Spotify data was performed to understand the factors that influence the odds of skipping a soundtrack recommended by Spotify. It was found that soundtrack recommendations in a shuffled playlist are 34 percent less likely to be skipped. Also, a Spotify premium user is 17 percent less likely to skip the soundtrack in comparison to a non-premium user. It can be inferred from this observation that Spotify has a better understanding of its premium users. A soundtrack recommendation based upon personalized playlist is 26 percent less likely to be skipped than a song recommendation from an editorial playlist. This implies that the Spotify personalized playlist is more individualistic and does cater to the taste of a user. There are several audio features like instrumentalness and acousticness which also impacts the odds of skipping a recommended soundtrack. A soundtrack with high instrumentalness is less likely to be skipped. Lastly, soundtrack recommended from 2000-2010 is less likely to be skipped than soundtrack recommendation from the 1970s.

## 1. Introduction

This study aims to analyze the factors that impact the odds of skipping a recommended soundtrack from audio features and user preferences. The analysis is performed on the modified dataset, originally obtained from Spotify, Inc. for Spotify Skip Prediction Challenge 2019. Due to the enormity of the dataset and computational limitations, data was aggregated and the study was performed on per sound-track basis. In this study, user preferences like listening to a shuffled playlist, owning a Spotify premium account and the kind of playlists the users listen to were analyzed for understanding the skipping behavior. Also, the audio features of sound-tracks like acousticsness, instrumentalness, liveness were studied.

Section 2 describes the data engineering, transformation, and exploratory data analysis to understand how different factors contribute to the odds of skipping a sound-track. In Section 3, model building is described, and the final model is presented. The evaluation of the final model and the statistical significance of individual parameters are also discussed in Section 4. Section 5 presents the summarized results that were obtained from model selection and conclusions inferred from the relationship. The limitations of the model are described in the final section.

## 2. Data

There are two datasets - user session logs and soundtrack metadata. The user log of soundtracks with ground truth skip label (hereafter referred to as *session-logs* dataset) has 137 million user sessions collected over two months. The soundtracks metadata(hereafter referred to as *sound-track* dataset) has 3.7 million soundtracks information. The dataset includes general metadata (e.g. duration, us popularity estimate, release year) as well as an audio features for all of these tracks. The songs released from 1950 to 1969 were removed primarily from the dataset for reducing the problem size.

### Data Preparation and Cleaning

The final dataset preparation steps and list of all the variables and their descriptions are mentioned in *Appendix A* for detailed reference. The final dataset obtained has information about 737 million recommendations of

0.3 million soundtracks collected for 1 month.

This study aims to understand how likely the recommended soundtracks are likely to be skipped. Because of this reason, it is more plausible to remove the songs which have been recommended quite less. The soundtracks with total recommendation of less than 50 in a month are removed based upon mean and standard deviation of total recommendations.

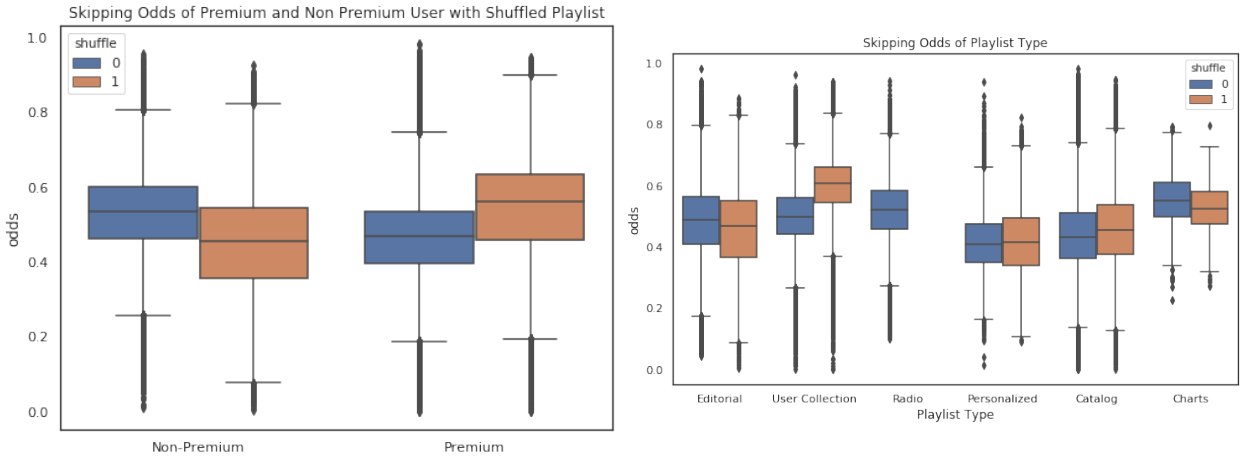
## Data Transformation

The release year of the song has been transformed into the decades of the soundtrack release years. This will provide a sensible interpretation of which decade songs are more likely to be skipped. The data was split into 70 percent training and 30 percent test dataset based upon the number of soundtracks using random sampling. In addition, all the continuous predictor variables (audio features) were centered around the mean.

## Exploratory Data Analysis

Out of 737 million recommendation of soundtracks, 50.8 percent of the recommendations are skipped. This implies that there is no requirement for oversampling or undersampling of data. The duration of the soundtracks has varying distribution for the odds of skipping. For soundtracks with duration less than the median, the skipping behavior is uniform. However, as the duration increases from the median, users are more likely to skip a song. (Ref *Appendix B*)

Among all the soundtracks recommendations, 31.09 percent of all the songs recommended across different sessions were listened by users with the shuffled playlist. From all these recommendations in the shuffled playlist, 54.8 percent of the songs were skipped by the user. In contrast, 49.03 percent of the soundtracks were skipped by the users without a shuffled playlist.




82 percent of the soundtrack recommendations were for the Spotify premium users. Among all these recommendations, premium users are slightly less likely to skip those songs in comparison to non-premium users. However, the difference seems to be marginal. The songs recommended to Spotify premium Users with shuffled playlist have a skipping rate of 55 percent, and without a shuffled playlist have a skipping rate of 48 percent. In contrast, songs recommended to Spotify non-premium Users with shuffled playlist have a 45 percent skipping rate and without shuffled playlist have a 52 percent skipping rate. Spotify non-premium users with shuffled playlist are least likely to skip and Spotify premium Users with shuffled playlist are more likely to skip.

There are 6 categories of soundtrack playlists for users to choose from. 40 percent of all the songs were recommended in the User Collection Playlist (e.g. a friend's playlist), 22 percent in Spotify Catalog (regional playlists), 14 percent in Spotify Radio and 19 percent in Spotify Editorial Playlist (from Spotify). There were 2 percent of the soundtracks recommendations in Personalized Playlist (personalized for every user) and 1.6 percent in the Spotify Charts (playlists based upon global or US popularity). The soundtracks recommended

for Personalized Playlist and Spotify Catalog had the least skipping rate. The highest skipping rate was observed among Spotify Charts and Spotify Radio.

The soundtracks recommendations in shuffled User Collection and Charts have 59 percent and 57 percent skipping rate respectively. The song recommendations in a non-shuffled Personalized Playlist have the least skipping rate of 42 percent, followed by non-shuffled Spotify Catalog. This trend is also observed in shuffled Personalized Playlist and shuffled Spotify Catalog.

The soundtrack recommendations across different sessions have 87.7 percent of the soundtracks from 2010 and beyond. There is about 6 percent of song recommendations that were released between 2000 and 2009. 3 percent of the recommendations are from the 1990s, 1.5 percent of soundtracks are from the 1970s and only 0.9 percent of the songs are from the 1980s. Among this distribution of song recommendations across decades, the likelihood of skipping is very similar all throughout at approximately 51-53 percent. (Ref *Appendix B*)

The US popularity estimate of the soundtracks is greater than 90 (out of 100) for all the soundtracks recommended in the dataset. It is observed that the soundtracks with lower popularity estimates (towards 90) have higher chances of getting skipped. However, as the popularity estimates increase, the skipping rate is uniform without any significant pattern. (Ref *Appendix B*)

The audio features of the recommended soundtracks have a uniform distribution of the skipping rate. However, specifically for 'Loudness', the soundtracks with relatively higher decibel value (-60 Db) tend to have a higher likelihood of being skipped. As the loudness converges to normal hearing (0 Db), the likelihood of skipping is uniformly distributed. In contrast, the lower the 'flatness', the lower the chances of skipping that song. However, similar to other features, as the flatness increases (converges to 1) the likelihood of skipping the song is uniformly distributed. (Ref *Appendix B*)

# 3. Model (Building and Assessment)

For model building, first, all the predictor variables were used and all of these variables were found to be statistically significant. Next, with AIC and BIC as the criteria- forward, backward and stepwise functions were used to generate models. The lower bound was set with predictor variables 'premium', 'shuffle', 'playlist_type', 'release_decade' and 'duration' for backward and stepwise selection. The upper bound was set with all the predictor variables about the user session preferences and audio features. Surprisingly, the final model obtained from this step was the same as the upper bound model.

Since all the predictor variables were found to be statistically significant, a check on multicollinearity was performed (Ref *Appendix C*). A correlation matrix was established and it was found that the variable beat_strength was highly correlated with danceability and bounciness. The variable beat_strength was moderately correlated with mechanism, speechiness, and valence. Also, the variable organism was moderately correlated with acousticness. The variable energy was highly correlated with loudness and moderately correlated with tempo. From all these correlation, the predictor variables acousticness, mechanism, speechiness instrumentalness, flatness, liveness, loudness, and valence were included in the further steps for model building. The model obtained from these variables had a drop in the AIC and BIC score.

The Interaction of premium flag with shuffle flag as well as shuffle flag with playlist type was subsequently added one by one on top of the model from the previous step. The two interactions were tested via a change in the deviance test and both of the interactions were found to be statistically significant. Model statistics of Accuracy, Sensitivity, Precision, and Specificity were also used to validate and compare each model across different stages. It was shown that by incorporating the two interaction terms, there was a slight increase in accuracy, with a substantial decrease in the AIC and BIC score. Hence, it was chosen as the final model.

To further validate the model, stratified cross-validation was performed on the training dataset with equal distribution of soundtracks based upon release decades on the training data. It was observed that for all the folds (k=10), the value of Accuracy, Sensitivity, Precision, and Specificity were very similar. This validated that the model was not overfitting or underfitting because of a skewed dataset wrt release decade.

In addition, binned plots were plotted for the continuous variables - acousticness, mechanism, speechiness

instrumentalness, flatness, liveness, loudness, and valence (*Appendix D*). The residuals were observed to be randomly distributed and quite a few outliers were observed. The confusion matrix and statistics of the model assessment on the training dataset are mentioned in *Appendix E*.

# 4. Result

The final model can be summarized as below:

$$logit(Pr[skip_i = 1]) = \beta_0 + \beta_1 Premium_i + \beta_2 Shuffle_i + \beta_{3,j} ReleaseYears[1980s, 1990s, 2000s, 2010s]$$

$$+\beta_{4,k} PlaylistType[User, Radio, Personalized, Catalog, Charts]_i + \beta_5 USPopularityEstimate_i + \beta_6 Acousticness_i +$$

$$\beta_7 Instrumentalness_i + \beta_8 Flatness_i + \beta_9 Loudness_i + \beta_{10} Mechanism + \beta_{11} Speechiness_i + \beta_{12} Liveness +$$

$$\beta_{13} Valence_i + \beta_{14,j} PlaylistType_i : Shuffle_i + \beta_{15} Premium_i : Shuffle_i + \epsilon_i$$

The statistical summary about the model is presented below:

| | Estimate | Exp(Coefficients) | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|---|
| (Intercept) | 1.202e-01 | 1.1107325 | 7.731e-04 | 155.48 | <2e-16 |
| premium | -1.947e-01 | 0.8289314 | 2.625e-04 | -741.65 | <2e-16 |
| release_decade2000s | 1.220e-01 | 1.1449522 | 1.559e-03 | 78.30 | <2e-16 |
| release_decade2010s | 3.914e-02 | 1.0504155 | 7.285e-04 | 53.72 | <2e-16 |
| release_decade1980s | 4.036e-02 | 1.0515494 | 1.111e-03 | 36.32 | <2e-16 |
| release_decade1990s | 9.025e-02 | 1.0963605 | 8.419e-04 | 107.19 | <2e-16 |
| shuffle | -4.865e-01 | 0.6186537 | 6.145e-04 | -791.69 | <2e-16 |
| playlist_type(User Collection) | -7.226e-02 | 0.9322346 | 3.038e-04 | -237.83 | <2e-16 |
| playlist_type(Radio) | 1.827e-01 | 1.2034491 | 3.390e-04 | 538.96 | <2e-16 |
| playlist_type(Personalized) | -2.917e-01 | 0.7498779 | 8.187e-04 | -356.31 | <2e-16 |
| playlist_type(Catalog) | -2.684e-01 | 0.7685149 | 3.277e-04 | -818.95 | <2e-16 |
| playlist_type(Charts) | 1.447e-01 | 1.1421823 | 9.183e-04 | 157.54 | <2e-16 |
| us_popularity_estimate | -1.796e-02 | 0.9767321 | 1.158e-04 | -155.03 | <2e-16 |
| duration | 3.945e-04 | 1.0004184 | 1.573e-06 | 250.78 | <2e-16 |
| acousticness | -1.051e-01 | 0.8987610 | 4.299e-04 | -244.37 | <2e-16 |
| instrumentalness | -3.398e-01 | 0.7261302 | 7.400e-04 | -459.18 | <2e-16 |
| flatness | -4.686e-02 | 0.9418855 | 2.639e-03 | -17.75 | <2e-16 |
| loudness | -3.474e-03 | 0.9975588 | 3.543e-05 | -98.05 | <2e-16 |
| mechanism | -7.904e-02 | 0.9105556 | 5.050e-04 | -156.50 | <2e-16 |
| speechiness | 1.431e-01 | 1.0949035 | 7.099e-04 | 201.60 | <2e-16 |
| liveness | 6.922e-02 | 1.0616221 | 6.194e-04 | 111.76 | <2e-16 |
| valence | 7.597e-03 | 1.0218163 | 4.471e-04 | 16.99 | <2e-16 |
| shuffle:playlist_type(User Collection) | 4.399e-01 | 1.5416113 | 4.845e-04 | 907.88 | <2e-16 |
| shuffle:playlist_type(Personalized) | -2.573e-02 | 0.9694798 | 1.336e-03 | -19.26 | <2e-16 |
| shuffle:playlist_type(Catalog) | 7.351e-02 | 1.0731425 | 5.956e-04 | 123.43 | <2e-16 |
| shuffle:playlist_type(Charts) | 1.002e-01 | 1.1268418 | 1.434e-03 | 69.86 | <2e-16 |
| premium:shuffle | 5.799e-01 | 1.7766115 | 5.545e-04 | 1045.82 | <2e-16 |

Table 1: Coefficients of Logistic Regression

The final model was definite in predicting the odds of skipping with an Accuracy, Sensitivity, Precision, and Specificity of 93.97%, 94.56%, 93.72%, and 94.23% respectively on the test dataset. The confusion matrix constructed from this model for the test data is mentioned in *Appendix E*.

The baseline of the model for inference is a soundtrack from the 1970s in a non-shuffled editorial playlist of a Spotify non-premium user with mean US popularity estimate and mean duration. This soundtrack has all the song audio features namely acousticness, mechanism, speechiness, instrumentalness, flatness, liveness, loudness and valence values at their respective mean from the dataset. This baseline soundtrack is 11.07 percent likely to be skipped by the user.

If the baseline soundtrack is recommended in a shuffled playlist, it is 38.2 percent less likely to be skipped than a non-shuffled playlist. Similarly, if the baseline soundtrack is recommended to a Spotify premium user, it is 17.1 percent less likely to be skipped than being recommended to a Spotify non-premium user. This validates our findings from EDA. If the baseline sound-track is recommended in a Spotify Radio playlist, it

is 20.3 percent more likely to be skipped. It is 14.2 percent more likely to be skipped if the soundtrack is recommended in Spotify Charts. In addition, if the baseline soundtrack is recommended in a Personalized Playlist or Spotify Catalog, it is 25 percent or 24.3 percent less likely to be skipped respectively.

Interaction effects of the shuffled playlist with a premium user as well as a shuffled playlist with playlist type provide further insights. For instance, if the baseline soundtrack is being recommended to a Spotify premium user with a shuffled playlist, it is 77 percent more likely to be skipped. Similarly, if the baseline track is recommended in a shuffled User Collection Playlist, it is 54.1 percent more likely to be skipped. This validates our findings in the EDA that recommendations in the shuffled User Collection Playlist are likely to be skipped the most. Also, if the baseline soundtrack is recommended in a shuffled Personalized Playlist, it is very unlikely (less than 3.1 percent) that the soundtrack will be skipped.

If a song with the same audio features as the baseline track but released in 2000s is recommended, it is 14.4 percent more likely to be skipped, keeping all other factors constant. A recommendation for a similar soundtrack with release year in the 1990s is 9.6 percent more likely to be skipped. If the soundtrack recommended has an increase in the US popularity estimate from the baseline by a factor of 1 from the mean, it is 2.4 percent less likely to be skipped. Also, if the duration of the baseline sound-track increases by 1 minute (60 seconds), the recommended soundtrack will be 2.5 percent more likely to be skipped.

If the instrumentalness of the recommended baseline soundtrack increases by 0.1 from the mean instrumentalness, the odds of skipping will decrease by 2.73 percent. Similarly, the odds of skipping will decrease by 0.1 percent, if the acousticness of the baseline recommended soundtrack increases by 0.1 from the mean.

# 5. Conclusion

By using a logistic regression model, we observe that several factors influence the skipping of a recommended song. The most important factor that affects the skipping behavior is if the soundtrack recommendation was in a shuffled playlist. A user listening to a shuffled playlist is 38.2 percent less likely to skip a recommended song. A Spotify premium user is 17.1 percent less likely to skip a recommended song than a non-premium user. A Spotify premium user with a shuffled playlist is 77 percent more likely to skip a recommended song than a Spotify premium user with a non-shuffled playlist. This implies that Spotify has a better understanding of the music taste of their premium user base, therefore they are less likely to skip. Spotify claims that the Personalized Playlist is catered for each individual and in this playlist a user is 26 percent less likely to skip a recommended song. This indicates that almost 1 out of every 4 soundtracks recommended to the user in the Personalized Playlist will be skipped. Furthermore, a user is 54.1 percent more likely to skip a recommended song from some other user's shuffled playlist (perhaps a friend).

The soundtracks with high instrumentalness are less likely to be skipped. This trend makes sense because a majority of people listening to instrumental music tend to tune in the music in the background and keep the music device aside. Hence, instrumental music is less likely to be skipped. Other audio features also impact the odds of skipping a song but it is marginally low.

# 6. Limitations

There were several limitations in the study which might indicate that the soundtrack skipping pattern might not be accurate. First, the data is skewed for several predictor variables, especially for playlist type and Spotify premium users. Furthermore, several of the audio features like US popularity is skewed towards the higher end and instrumentalness is skewed towards the lower end. This might result in overfitting or underfitting the model. The model provides pretty good accuracy and other scores for model evaluation metrics. The ROC curve could not be plotted as the data was aggregated and it was impossible to collapse the data because of limited computational resources. Because of this, no optimal threshold value was found. Lastly, the binned plots of residuals with continuous variables represent a substantial number of outliers. However, nothing much could be done in this regard to eliminating those outliers.

# 7. Appendix

## Appendix A: Data Preparation

Initially, the daily session-log files are merged and aggregated together with composite key as track_id, skip flag, premium flag, shuffle flag, and playlist type. Then, two separate data frames are created based upon skip flag for obtaining the skipped and no skipped count per soundtrack. An outer join is performed on the skipped and no-skipped count data frame with the same composite key to generate a single data frame. This resulting data frame is joined with the sound-track frame, to get the final dataset with skipped and not skipped count per soundtrack.

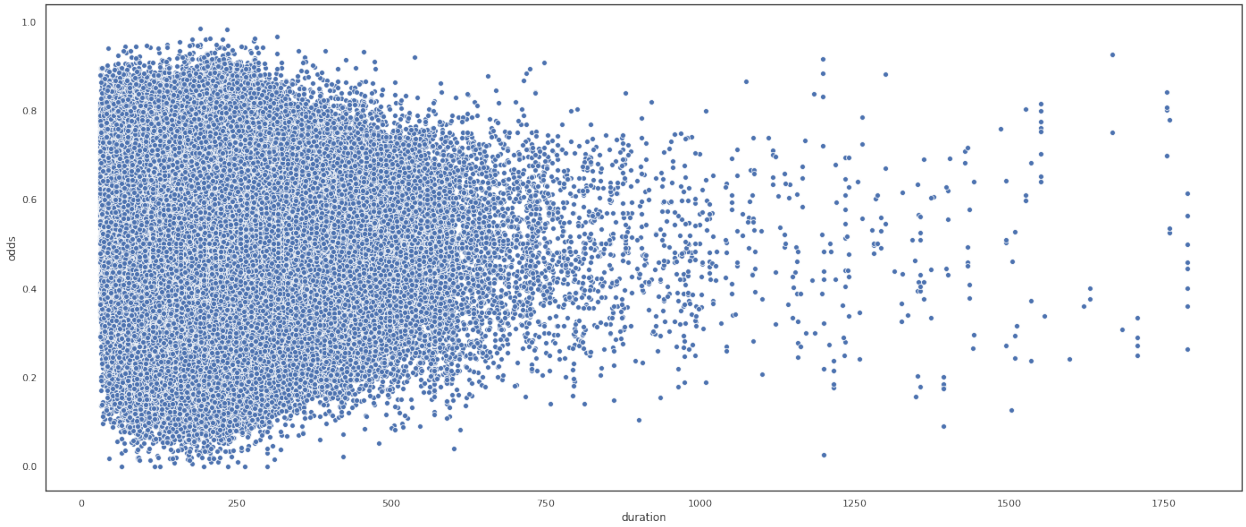
### Variable Description

| Variable Name | Type | Description |
|---|---|---|
| track_id | string | hashed track id |
| premium | bool | Spotify premium user Yes(1) or No(0) |
| shuffle | bool | Shuffled Playlist Yes(1) or No(0) |
| playlist_type | categorical | For description of type of playlist refer to https://artists.spotify.com/guide/playlists. editorial_playlist=0, user_collection=1, radio=2,personalized_playlist=3, catalog=4, charts=5 |
| duration | numeric | duration of soundtrack |
| release_year | numeric | release year of soundtrack |
| us_popularity_estimate | numeric | US popularity Estimate |
| acousticness | numeric | https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/ |
| beat_strength | numeric | Same as accousticness |
| bounciness | numeric | Same as accousticness |
| danceability | numeric | Same as accousticness |
| energy | numeric | Same as accousticness |
| flatness | numeric | Same as accousticness |
| instrumentalness | numeric | Same as accousticness |
| liveness | numeric | Same as accousticness |
| loudness | numeric | Same as accousticness |
| mechanism | numeric | Same as accousticness |
| organism | numeric | Same as accousticness |
| speechiness | numeric | Same as accousticness |
| tempo | numeric | Same as accousticness |
| valence | numeric | Same as accousticness |
| skipped_count | numeric | Number of users who skipped the soundtrack |
| not_skipped_count | numeric | Number of users who did not skipped the soundtrack |
| total_count | numeric | Number of users who skipped + who did not skip the soundtrack |
| release_decade | categorical | Release decades from 1970s, 1980s, 1990s, 2000s, 2010s |

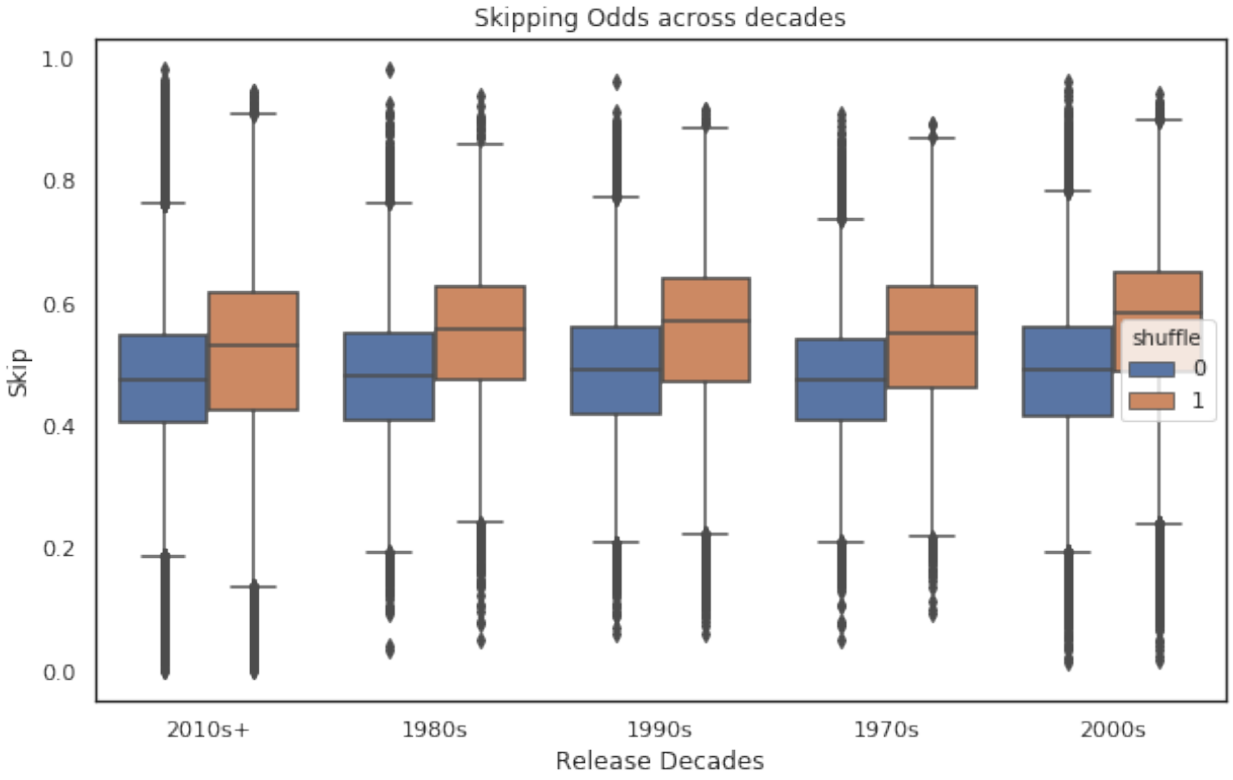
Table 2: Coefficients of Logistic Regression

## Appendix B: EDA Plots

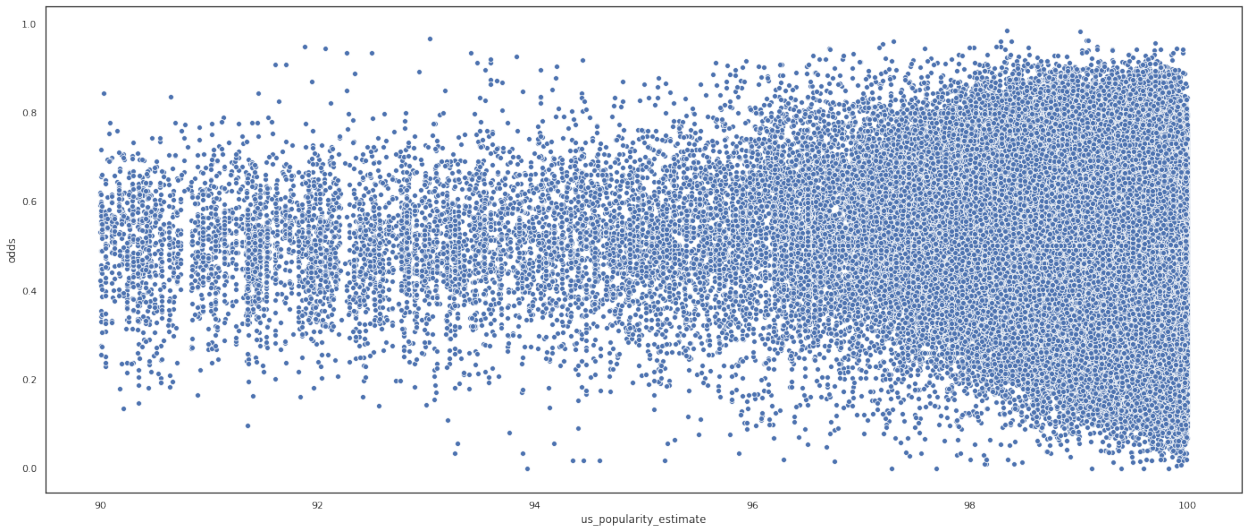
### Soundtrack Duration wrt Skipping ratio



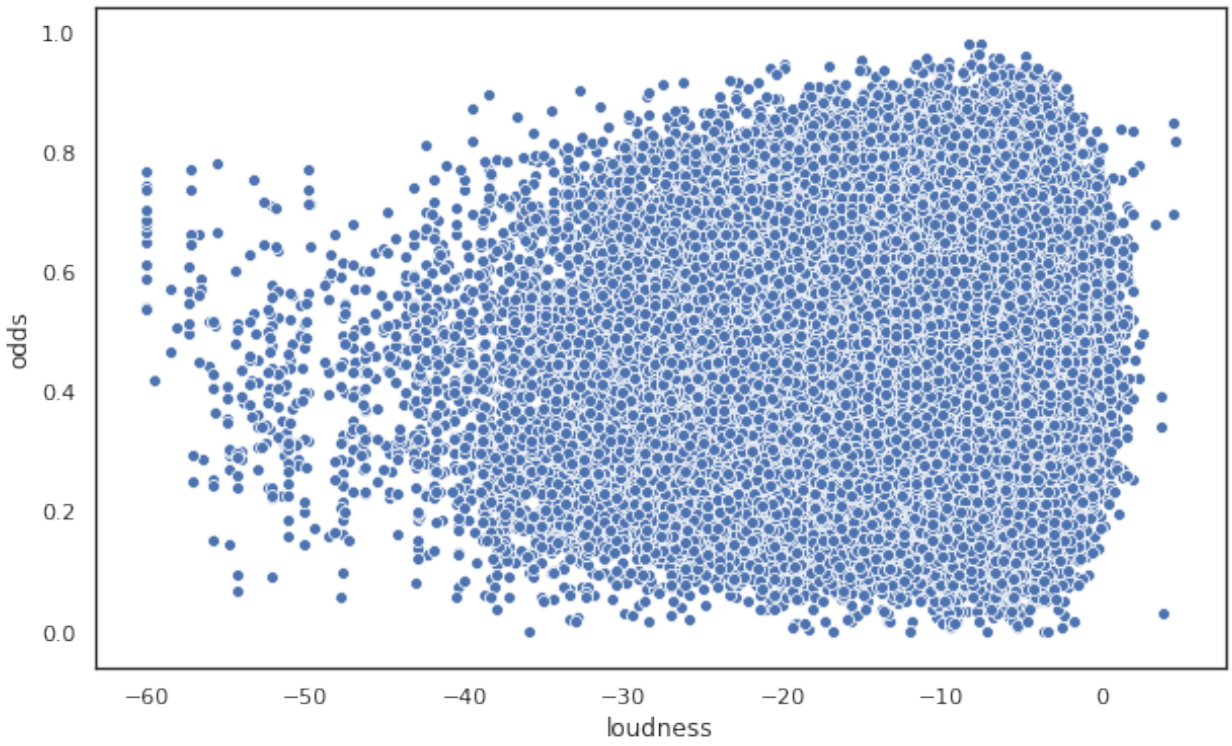

### Release Decade wrt Skipping ratio



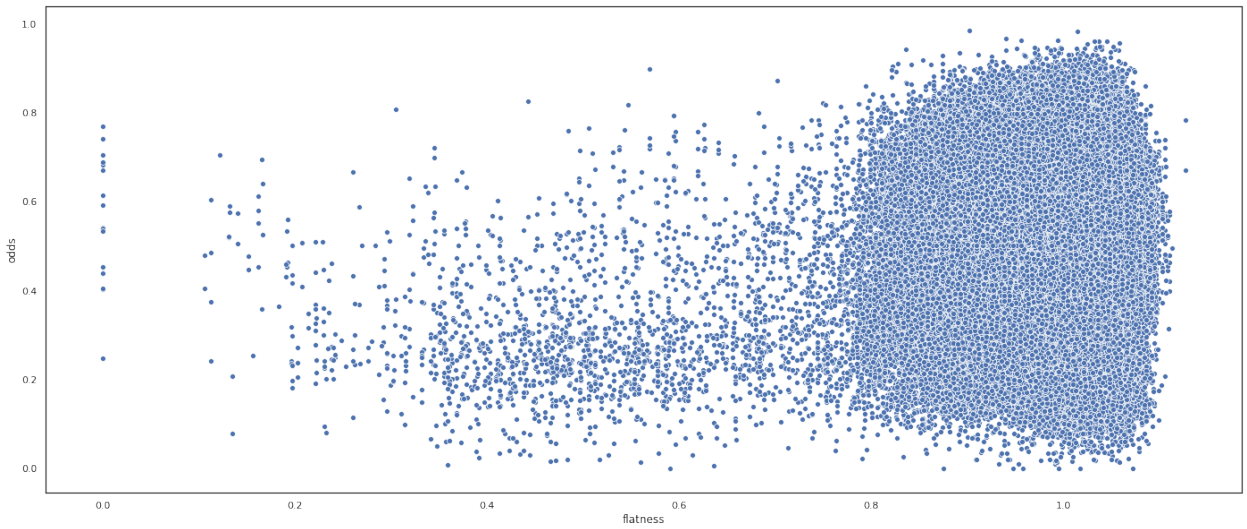
**US popularity Estimate wrt Skipping ratio**



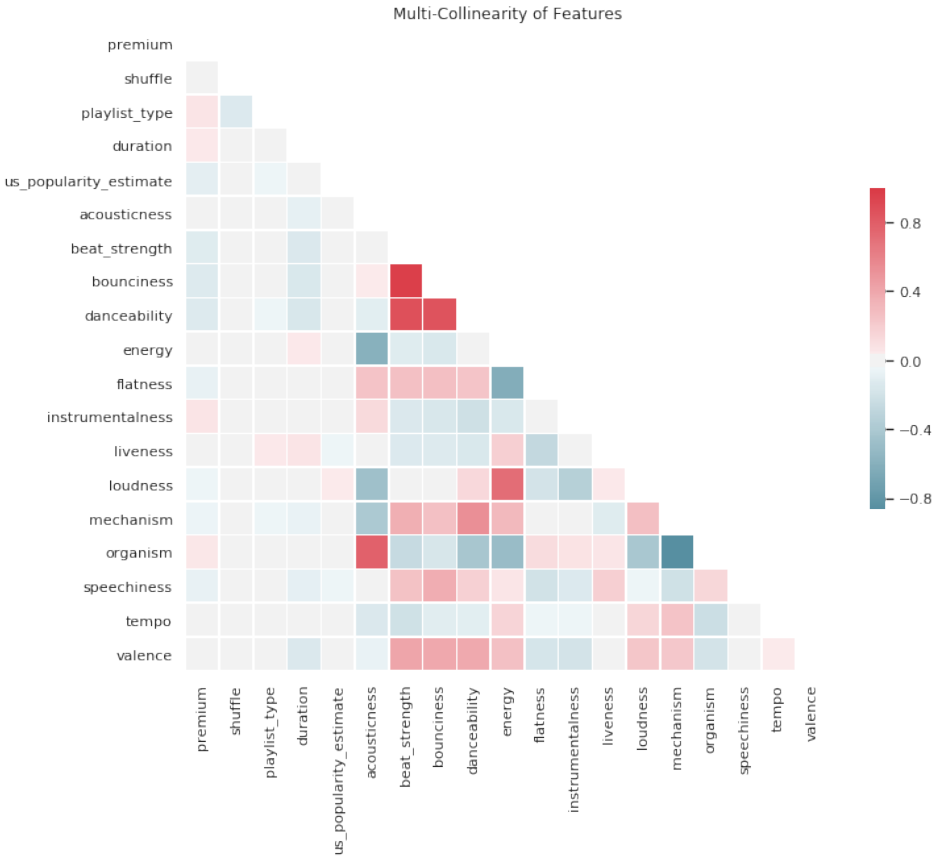

**Loudness wrt Skipping ratio**



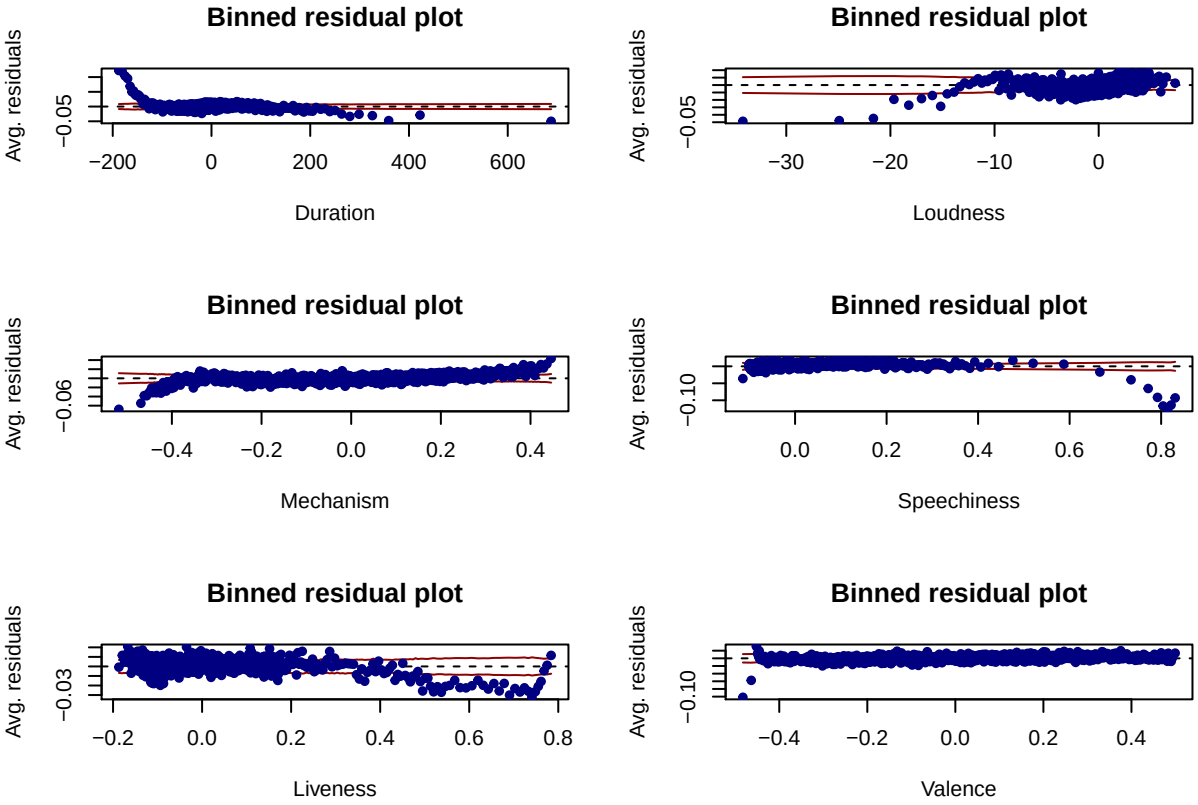
### Flatness wrt Skipping ratio




## Appendix C: Test for Multicolinearity



## Appendix D: Binned Residual Plots




## Appendix E: Training Dataset metrics

| | Positive Prediction | Negative Prediction |
|---|---|---|
| Positive Classification | 262742789 | 16957641 |
| Negative Classification | 17360944 | 255930657 |

Table 3: Confusion Matrix

| Parameter | Value |
|---|---|
| Accuracy | 93.79% |
| Specificity | 93.64% |
| Sensitivity | 93.8% |
| Precision | 93.9% |

Table 4: Model Strength Params on training data

## Appendix F: Testing Dataset metrics

| | Positive Prediction | Negative Prediction |
|---|---|---|
| Positive Classification | 111842909 | 7481796 |
| Negative Classification | 6427794 | 105009702 |

Table 5: Confusion Matrix