Spotify – Recommendations to Reduce Song Skipping

Marketing Models
Columbia University in the City of New York
Isha Gupta
isha.gupta@columbia.edu | **LinkedIn**

Abstract: *Spotify is one of the top consumer-preferred music streaming platforms. They are known for providing recommendations to the consumers with the latest, relevant, and consumer-preferred songs. Although, as the data reveals, a lot of users skip the songs while being recommended in a playlist. Some skip early, some late while some not at all. Being data scientists, we dived deeper into knowing the reasons behind a song being skipped and provide possible recommendations for Spotify to make their users love songs more and skip less.*

1. **Goal of the Project:** Build and estimate a model to predict the user's skipping behaviour and derive managerial insights for Spotify, specifically, answering the question - how can Spotify reduce user-skipping!

2. **Data Description:**

The provided dataset (a simplified subset of Spotify AI challenge) consisted of 9 acoustic features of songs and 2 target variables – 'early skip' and 'late skip'. The meaning of the features is as follows:

1. Duration: Duration or length of the song in seconds
2. Acousticness**:** The confidence measure of how natural acoustic sounds it consists of. Lower acousticness means higher electric sounds. 1.0 represents high confidence the track is acoustic
3. Danceability**:** The extent of 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 1.0 means most danceable
4. Energy: A measure from 0.0 to 1.0 representing a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy
5. Instrumentalness**:** Measure of 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". 1.0 indicates greater likelihood that the track contains no vocal content
6. Liveliness**:** Detection of the presence of an audience in the recording. Higher liveliness representing increased probability that the track was performed live
7. Loudness**:** Overall loudness of the track in decibels (dB). is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db
8. Tempo: It is the speed or pace of a given piece derived directly from the average beat duration in bpm (beats per minute)
9. Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track
10. Early: If song was skipped early or not. 1 indicating early skip while 0 could mean either a late skip or no skip
11. Late: If song was skipped late or not. 1 indicating late skip while 0 indicating no skip.

Note: '1' in Early column had '1' in late column as well

3. **Feature Engineering:** To account for redundancy in 'Early' and 'Late' column. A new column was created named as 'Skip_speed' by adding both columns. '0' indicating 'No skip', '1' indicating 'late skip' and '2' indicating 'early skip'.

4. **Modelling:**

Firstly, the trained dataset was scaled and test set was scaled on the train's mean and standard deviation.

4.1 No-Skip – Late Skip – Early Skip Classifier | 3 Class Multi-class classifier

First aim was to find the best classifier with highest accuracy or least loss on the hold-out test set.

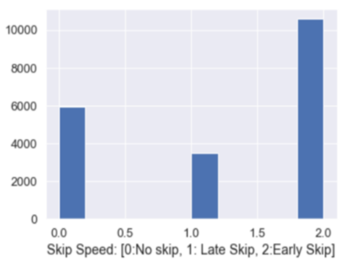


Figure 1 Frequencies of each class

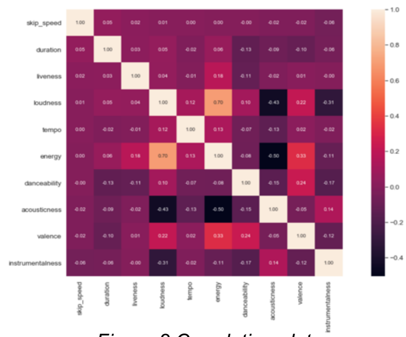


Figure 2 Correlation plot

	duration	acousticness	danceability	energy	instrumentalness	liveness	loudness	tempo	valence	early	late
count	20000.0000	20000.0000	20000.0000	20000.0000	20000.0000	20000.0000	20000.0000	20000.0000	20000.0000	20000.0000	20000.0000
mean	219.4678	0.2253	0.6518	0.6346	0.0350	0.1895	-7.0106	122.4435	0.4752	0.5289	0.7042
std	61.8584	0.2515	0.1572	0.1887	0.1522	0.1528	3.2282	29.8700	0.2343	0.4992	0.4564
min	31.3981	0.0000	0.0000	0.0002	0.0000	0.0200	-51.1770	0.0000	0.0000	0.0000	0.0000
25%	185.8908	0.0300	0.5469	0.5122	0.0000	0.0984	-8.3710	97.0430	0.2877	0.0000	0.0000
50%	214.0467	0.1212	0.6687	0.6438	0.0000	0.1266	-6.3895	122.0990	0.4599	1.0000	1.0000
75%	246.3800	0.3446	0.7673	0.7795	0.0001	0.2358	-4.9178	142.9705	0.6506	1.0000	1.0000
max	875.3066	0.9958	0.9876	0.9998	0.9923	0.9919	0.1750	216.4620	0.9873	1.0000	1.0000

Figure 3 Summary Statistics

1. Dataset was imbalanced.
2. No strong correlation with skip speed was observed for any of the song features. Although, loudness and acousticness as well as energy and acousticness were highly negatively correlated.

4.2 Metrics for model comparison

Accuracy: It is the ratio of number of correct predictions over total predictions made. Accuracy is one of the most commonly chosen metrics of a classifier

Log loss *(Logarithmic loss)* or Categorical Cross Entropy loss : It is used as one of the best metrices to evaluate a classifier. Unlike acccuracy which just measures 'yes's or 'no's, log loss takes into account the uncertainty of the prediction based on how much it varies from the actual label. Thus, giving a more nuanced view into the performance of the model.

$$Log\ Loss\ =\ \sum - y.log\ log\ (p)\ - (1-y).log\ (1-p)$$

In this paper, 5 models will be compared with another on these 2 metrices. Accuracy is also chosen, as trees inherently predict the class directly from the trained splits, thus, Accuracy makes more sense for the trees.

4.3 Multinomial Logistic Regression

Multinomial Logistic Regression : It is the simplest Machine Learning Model and provides high interpretabiliy of feature importances for every target class. Multinomial log loss on hold out test set was 0.99345.

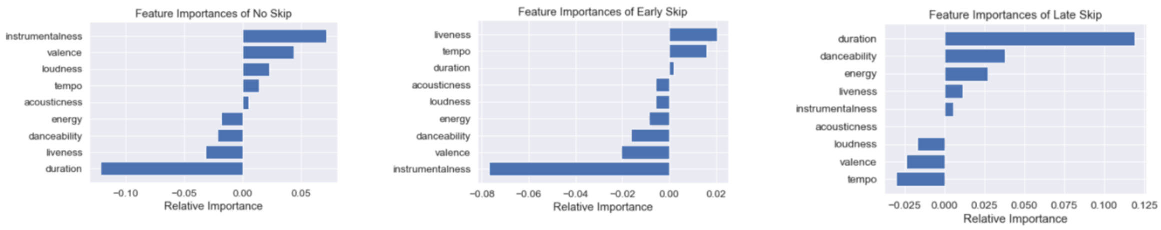


Figure 4. Feature importance for each class for multinomial regression

4.3.1 Interpretations:

- Lack of vocals (high instrumentalness) in song reduce skipping, especially early skipping significantly
- Positivity (high valence) in song makes the user listen song without skipping
- Long duration songs lead to high number of late skips indicating that user tend to skip the song after a certain time of listening
- Liveliness lead to early skipping of songs – implying that as soon as user identifies that song is not original or there is 'noise' of audience, user tends to skip the song.
- Loudness being positively correlated with 'No skip' and negatively correlated with both the 'early' and 'late' skips indicate that louder the song, lesser the chances of user to skip it
- Danceability is negatively correlated to both the 'no skip' and 'early skip' while highly positively correlated to 'late skip' implying that high danceability makes the user skip the song although late
- Energy is negatively correlated to both the 'no skip' and 'early skip' while highly positively correlated to 'late skip' implying that high energy songs makes the user skip the song although late
- High tempo in song lead to more early skips and lesser late skips
- Acousticness has relatively the least affect on user's skipping behaviour

4.4 **Decision and Gradient Boosting Trees:** Trees are one of the best interpretable Machine Learning models. For the 3 class multi-class classification using decision tree, it was tuned through cross – validation by finding the best values of `max_depth, ccp_alpha (complexity parameter), criterion` and `splitter`

Performance of Neural network and ensemble trees increased from 53% to 72% when changing from 3 class classifier to 2 class classifier. It showed that it is better and more reliable to first classify if the song will be skipped at all or not and make first level of recommendations and then conditioned on being skipped, creating another classifier if a song will be skipped early or late.

4.4.1 **Skip V/s No-skip Classifier**

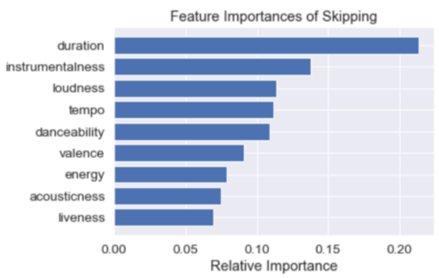


Figure 5. Feature importance for Skip – No Skip classifier: GradientBoosting (Ensemble of 100 decision trees)

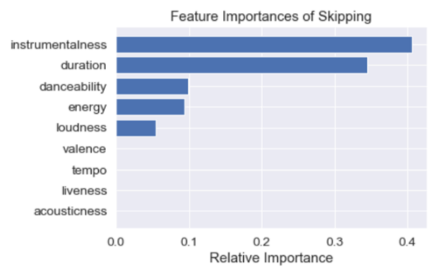


Figure 6. Feature importance for Skip – No Skip classifier: Decision Tree (single tree)

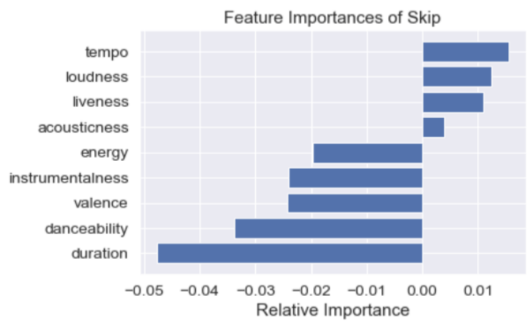


Figure 7. Feature importance for Skip – No Skip classifier: Logistic Regression- Elasticnet

Interpretation:

- Length of songs increases chances of skipping
- Lack of vocals (high instrumentalness) increases chances of skipping
- Loudness increases chances of skipping
- Danceability increase chances of skipping

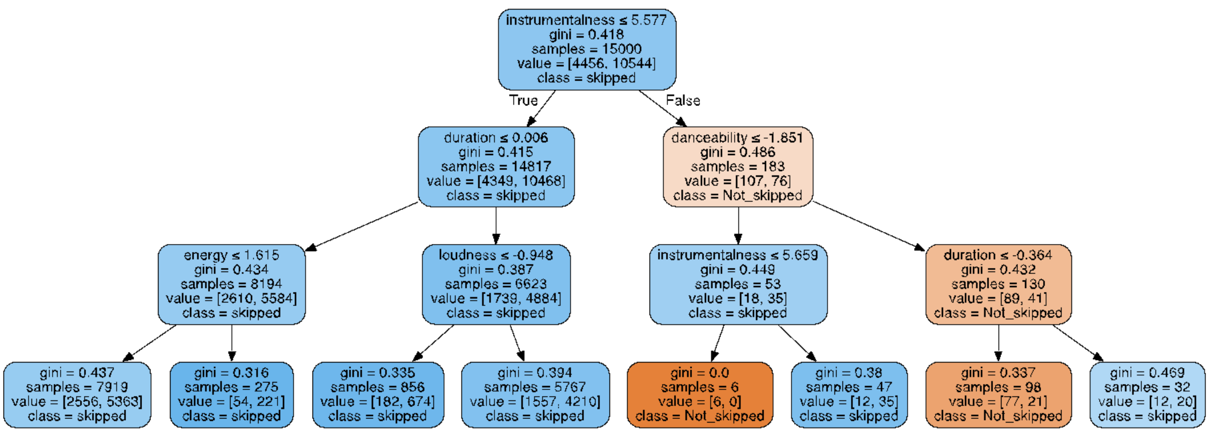


Figure 8. Skip – No Skip classifier Decision Tree

4.4.2 Early Skip V/s Late skip Classifier

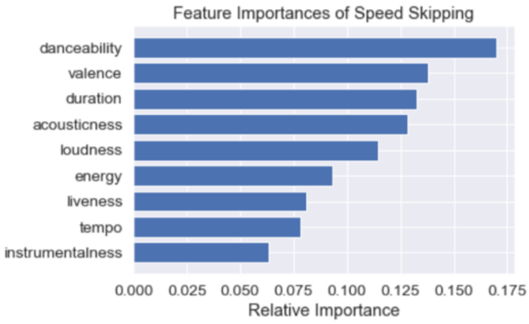


Figure 9. Feature importance for Early Skip – Late Skip classifier: GradientBoosting (Ensemble of 100 decision trees)

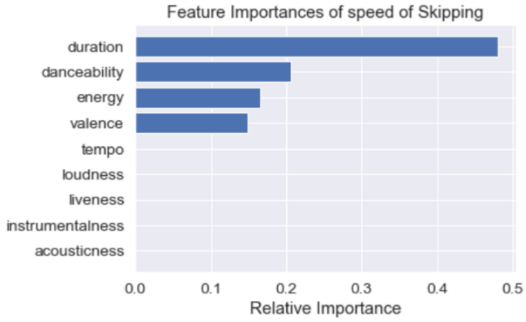


Figure 10. Feature importance for Early Skip – Late Skip classifier: Decision Tree (single tree)

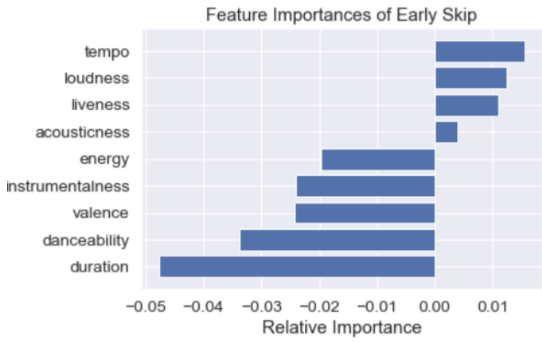


Figure 11. Feature importance for Early Skip – Late Skip classifier: Logistic Regression- Elasticnet

Interpretation:

- Length of songs increases speed of skipping
- Danceability increases speed of skipping
- Valence increases speed of skipping
- Energy (High intensity and activity) in songs increase the speed of skipping

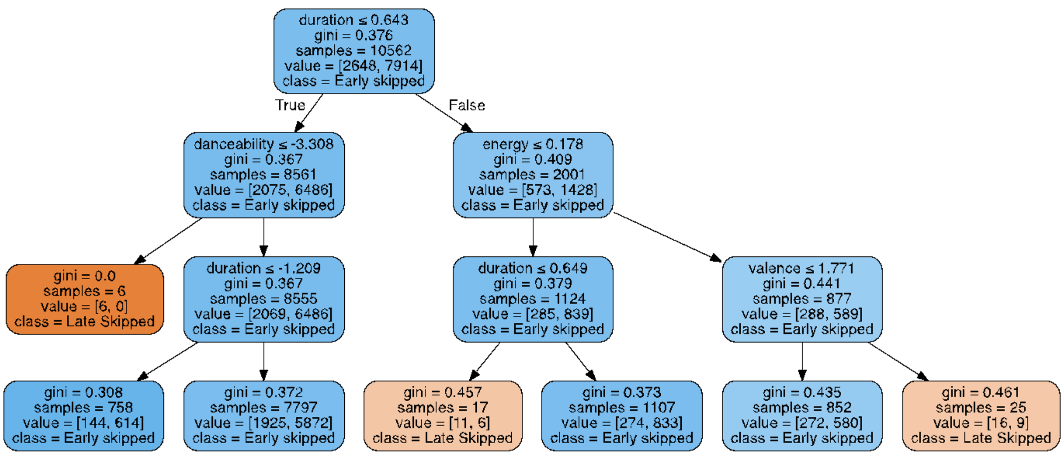


Figure 12. Early Skip – Late Skip classifier Decision Tree

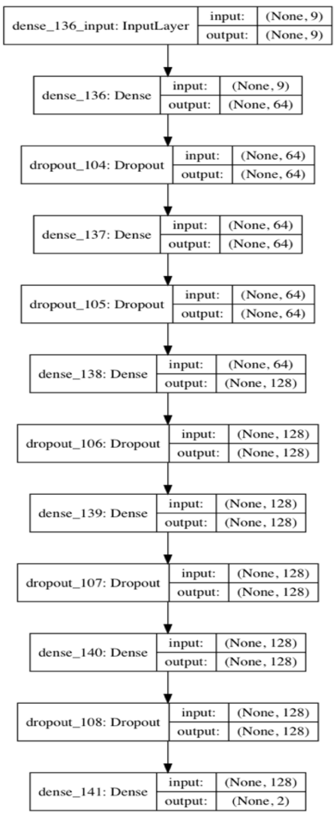
4.5 Modelling Summary:

Modelling Method	No skip: Early Skip : Late Skip classifier		Skip : No-skip classifier		Early Skip : Late Skip Classifier	
	Log – Loss	Accuracy	Log – Loss	Accuracy	Log – Loss	Accuracy
Multinomial Logistic Regression (Elasticnet)	0.99345	0.5386	0.99286	0.5382	0.55483	0.75660
SVM	0.99449	0.5392	0.99444	0.5402	0.55528	0.75604
Neural Network	0.99554	0.5382	0.59901	0.71146	0.55804	0.75660
Decision Tree	0.99591	0.5366	0.99522	0.5408	8.42644	0.75603
Gradient Boosting	0.99575	0.538	0.99518	0.5374	8.51472	0.75348

Above table shows the test accuracy (accuracy on the 5000 hold-out samples) of each classifier which was trained and tuned through cross-validation on 15000 training samples.
Results show that 3 class classifiers perform very poorly as compared to 2 separate classifiers. The reason is that 3 classes are not exactly equivalent. A song being skipped or not skipped is about if a listener loved it or hated it, while a song being skipped early or late is about how soon the listener dislikes or get bored of the song. The classes in a multiclass classifier are assumed to be equivalent or equidistant from each other which is not applicable here.

Hence, we moved to find best classifiers both for **'Skip-No Skip'** as well as for **'Early skip-Late Skip'.** Binary classifiers were found to not only produce much improved results but also provide better visualization and interpretation for making better business recommendations. Hence, our final predictor will first use Neural Network model to predict whether the given song will be skipped or not and then use the logistic regression to predict the speed of skipping (early or late skipping) conditioned on the song being skipped. Important thing to note here is that, in case of **Skip-No Skip'** classifier, neural network outperforms all other classifiers. Whereas in case of **Early skip-Late Skip'** classifier, all classifiers are improved to similar extent. Given the high computational power required by SVM, low interpretability of Neural Network and simplicity of logistic regression, we chose the latter to form 2nd layer of predictor.

Input song (9 features)



[Probability of Not skipping , Probability of Skipping]

Figure 13. Classifier Layer 1 : Neural Network

Input song (9 features)

$$\textbf{Probability}\,(\boldsymbol{early\ skipping}) = \beta_0 * duration + \beta_1 * acousticness + \beta_2 * danceability + \beta_3 * energy + \beta_4 * instrumentalness + \beta_5 * liveliness + \beta_6 * loudness + \beta_7 * tempo + \beta_8 * valence$$

B_0	-0.047573
B_1	0.003921
B_2	-0.033735
B_3	-0.019800
B_4	-0.024146
B_5	0.010877
B_6	0.012336
B_7	0.015453
B_8	-0.024296

p

Probability of early skipping = probability (early skipping | skipping)

Probability of late skipping = probability (late skipping | skipping)

Probability of early skipping = *p x* ***Probability of Skipping***

Probability of late skipping = *(1 – p) x* ***Probability of Skipping***

Figure 14. Classifier Layer 2 : Elasticnet Logistic Regression

5. **Business Recommendations To Spotify**

5.1 Recommendations to Reduce Skipping:

As decision trees are inherently classifiers, log-loss doesn't perform well as their metric as it makes them calculate probabilties. From the decision tree splits and feature importances above, following are our top recommendations for spotify:

- **Lower cap on instrumentalness:** Instrumentalness of the recommended song should be greater than 0.8808. This is clear from the 'Skip-no_skip' decision tree plot, which shows the scaled values of all variables. From the first split, it can be seen that if instrumentalness is below 0.8808 then the song is definitely skipped by the listener.

- **Instrumentalness and danceability interaction:** If danceability of a song is less than 0.362, then instrumentalness shouldn't exceed beyond 0.8932. It implies that if there aren't enough musical elements in the song including tempo, rhythm stability, beat strength, and overall regularity, then person tend to skip if even vocals are missing beyond a certain level. It makes logical sense as well, when a person doesn't find any interesting music, beats or rhythm then she/he should find lyrics to dwell in else person skips it.
  Therefore, a playlist can be created such that to maintain entertainment and reduce monotonicity, for instance a lyrical song is followed by an instrumental song of genre of user's preferences or mood or vice versa.

- **Danceability effect:** The danceability should be higher than 0.362 to ensure no skipping. It implies that if vocals are lacking beyond a certain level, then song should be highly danceable i.e. should have strong tempo, rhythmic stability, beat strength and overall regularity for the person to keep listening.

- **Duration effect:** If duration of song exceeds beyond 196.95 seconds then person will skip the song. It shows that after a certain period of time, person gets bored of listening the same song and has enjoyed every element of the song.
  Therefore, to avoid it, in the mash-ups and spotify curated mood playlists, especially those which listeners listen and choose to like or dislike, should not have complete songs rather only till certain length. So that, as soon as the person starts getting bored of listening one song, the next song is played automatically without person having to skip the song which reduces customer satisfaction.

5.2 Recommendations to Delay Skipping:

- **Duration and danceability interaction:** If duration of song is less than 182 seconds then danceability should be less than 0.134 to delay the skipping. In violation of this, person tends to skip song earlier. To keep person hooked for longer danceability should be less for relatively shorter songs. It makes intuitive sense as well. Danceable song are quickly chosen by listeners – 'to listen or to not listen' or rather 'to dance on or not' but other songs have to be listened for a relatively longer duration to make a decision. The reason is danceable songs can be liked or disliked merely by beats, vibes and rhythm while other songs have to be given time to judge through other elements.
  This implies that while recommending songs to user, recommended danceable songs should be of user's choice – either through previous history or previous interests while a safe 'experiment' can be done with non-danceable songs, as user would have to 'listen' to the new song to decide before quickly skipping.
- **Energy and duration effect:** If energy of a song is less than 0.756 then the duration should definitely be less than 262 seconds. This implies that if a song has low energy level and duration extends beyond a limit, then listener tend to skip the song early. In other words, low energy songs can hook a listener till a certain limit, beyond which user tend to skip it.
  This implies that while recommending new or out of the box latest low energy songs, then shorter duration of that song should be played, and rather than keeping the complete song in the playlist, next song should be played after 2nd or 3rd verse itself.
- **Valence and Energy effect:** If song has energy levels above 0.756, then it should have high valence i.e. positivity too above at least 0.887. This implies that high energy songs with negativity in them i.e. low valence, tend to be skipped by listner earlier. Hence, among the generalized recommendation songs if the energy level of song is high they should be imparting a positive emotion rather than a negative one.

5.3 Future Recommendations

- **Confounding Variables:** As this data consisted of only acoustic features of songs, there could possibly be other confounding or hidden variables such as listeners' demographics, geographical region, cultural background etc. that could highly affect the skipping behaviour as it is a personal choice and same song can be loved by one listener and hated by another. Building separate models for separate age groups or geographical regions might perform better due to change in tastes with age and regions.
  Also, these are only 20000 samples of songs which might be a biased representative of the dataset such as songs belonging to a particular genre, a particular set of artists, popularity level etc. Therefore, for a more robust model building, it should be ensured that only acoustic features are main differentiating variables among these songs and rest all differences are eliminated by proper random sampling from the actual dataset.
- **Iterative Improvement:** Based on recorded real-time customer satisfaction score after implementation of recommendations, a next set of strategies can be implemented for those customers which are hard to please. To evaluate the impact of implemented strategies, it should be essential to observe, how much skipping is reduced for different type of listeners and then perform next change of strategies for different listener segments.

$$\boldsymbol{CS^{*}\,(x) = x\,*\,\beta + \epsilon}$$

Where x are session and listener descriptors such as time, day, age, gender, country etc.

- **Optimal order of songs in recommended playlists:** If arrangement of songs in the recommended playlists is such that next song is more 'likeable' or 'more user's favorite' then users will develop an inherent tendency of skipping a song to listen the next 'better' song. On the other hand, if every next song in line is worse or less favorite then user might simply drop off. Therefore, through A/B testing or random test-control trials – the right re-ordering of songs need to be decided to maximize profit from the newly recommended playlist.
- **Metrics of Virality:** It should be useful to analyze that at a particular time – when certain songs are viral in a particular region or among a certain age groups or for certain genre of songs, then are users skipping 'till' that song is found or is the skipping behaviour usual. Such analysis could lead to better insights towards viral effect of songs and extent of virality among a certain set of listeners.