SPOTIFY SEQUENTIAL SKIP PREDICTION

(CLASSIFICATION PROBLEM)

PRESENTED BY:

PRAVESH RAIKWAR

NIHAR SHAH

FEATURES IN THE DATA

Table 1(user log)

```
0    session_id                      167880 non-null   object
1    session_position                167880 non-null   int64
2    session_length                  167880 non-null   int64
3    track_id_clean                  167880 non-null   object
4    skip_1                          167880 non-null   bool
5    skip_2                          167880 non-null   bool
6    skip_3                          167880 non-null   bool
7    not_skipped                     167880 non-null   bool
8    context_switch                  167880 non-null   int64
9    no_pause_before_play            167880 non-null   int64
10   short_pause_before_play         167880 non-null   int64
11   long_pause_before_play          167880 non-null   int64
12   hist_user_behavior_n_seekfwd    167880 non-null   int64
13   hist_user_behavior_n_seekback   167880 non-null   int64
14   hist_user_behavior_is_shuffle   167880 non-null   bool
15   hour_of_day                     167880 non-null   int64
16   date                            167880 non-null   object
17   premium                         167880 non-null   bool
18   context_type                    167880 non-null   object
19   hist_user_behavior_reason_start 167880 non-null   object
20   hist_user_behavior_reason_end   167880 non-null   object
```

Table 2 (song data)

```
0    track_id                50704 non-null   object
1    duration                50704 non-null   float64
2    release_year            50704 non-null   int64
3    us_popularity_estimate  50704 non-null   float64
4    acousticness            50704 non-null   float64
5    beat_strength           50704 non-null   float64
6    bounciness              50704 non-null   float64
7    danceability            50704 non-null   float64
8    dyn_range_mean          50704 non-null   float64
9    energy                  50704 non-null   float64
10   flatness                50704 non-null   float64
11   instrumentalness        50704 non-null   float64
12   key                     50704 non-null   int64
13   liveness                50704 non-null   float64
14   loudness                50704 non-null   float64
15   mechanism               50704 non-null   float64
16   mode                    50704 non-null   object
17   organism                50704 non-null   float64
18   speechiness             50704 non-null   float64
19   tempo                   50704 non-null   float64
20   time_signature          50704 non-null   int64
21   valence                 50704 non-null   float64
22   acoustic_vector_0       50704 non-null   float64
23   acoustic_vector_1       50704 non-null   float64
24   acoustic_vector_2       50704 non-null   float64
25   acoustic_vector_3       50704 non-null   float64
26   acoustic_vector_4       50704 non-null   float64
27   acoustic_vector_5       50704 non-null   float64
28   acoustic_vector_6       50704 non-null   float64
29   acoustic_vector_7       50704 non-null   float64
```
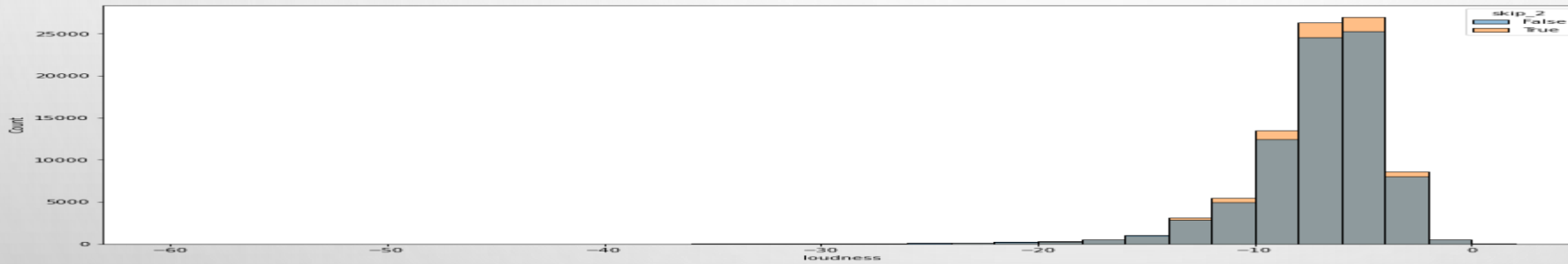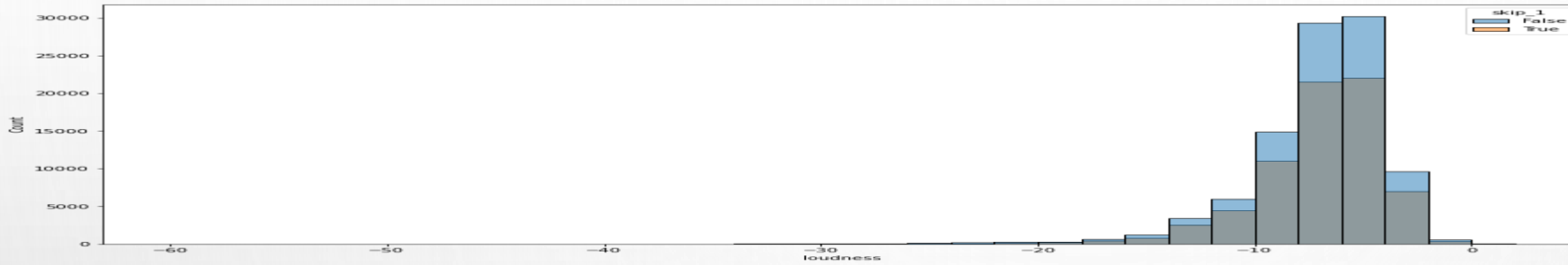
+

UNDERSTANDING THE DATA
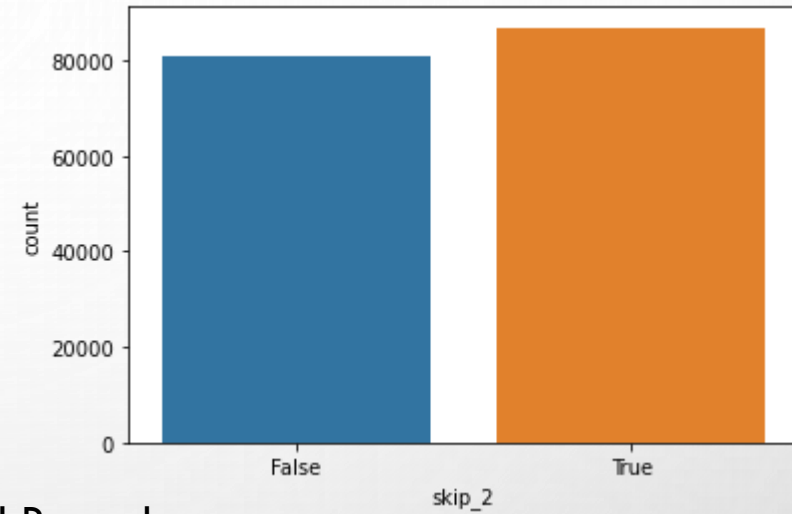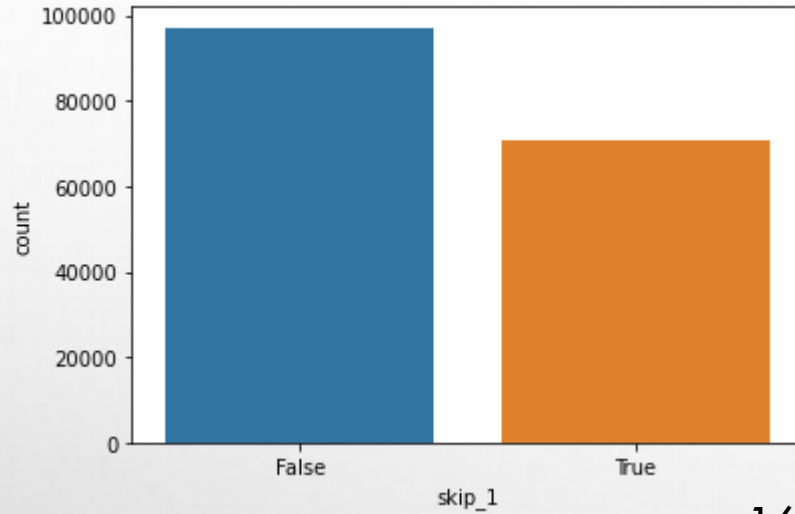
167880 Total Records
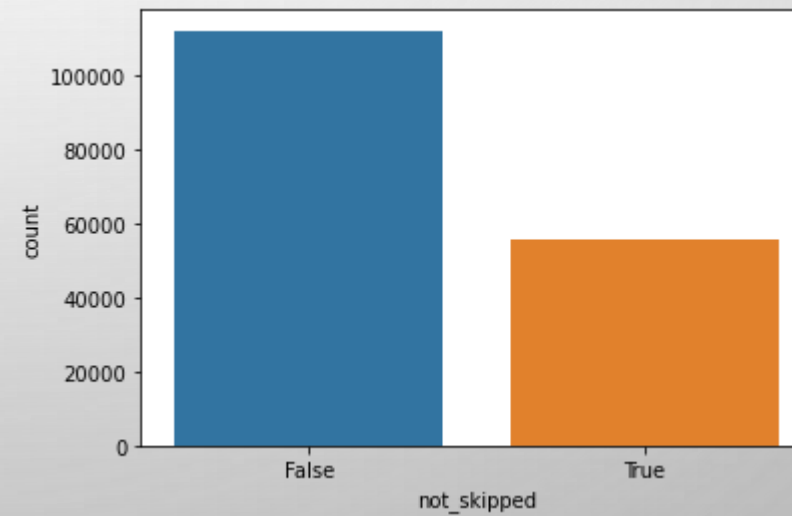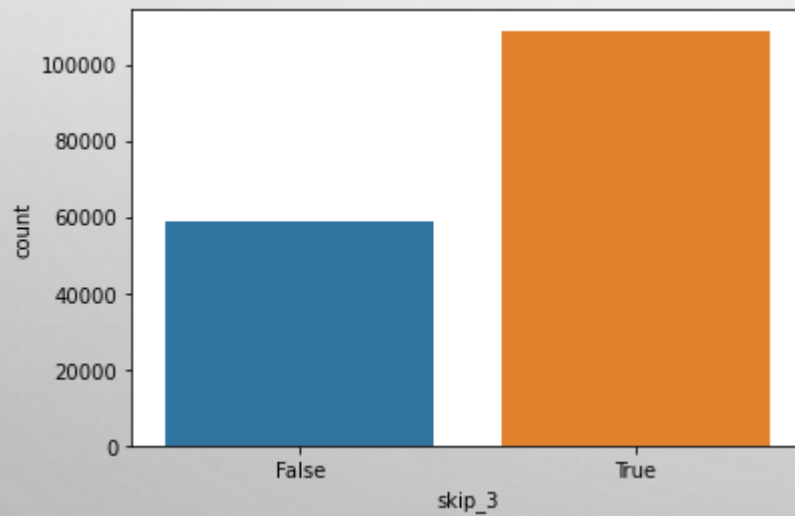
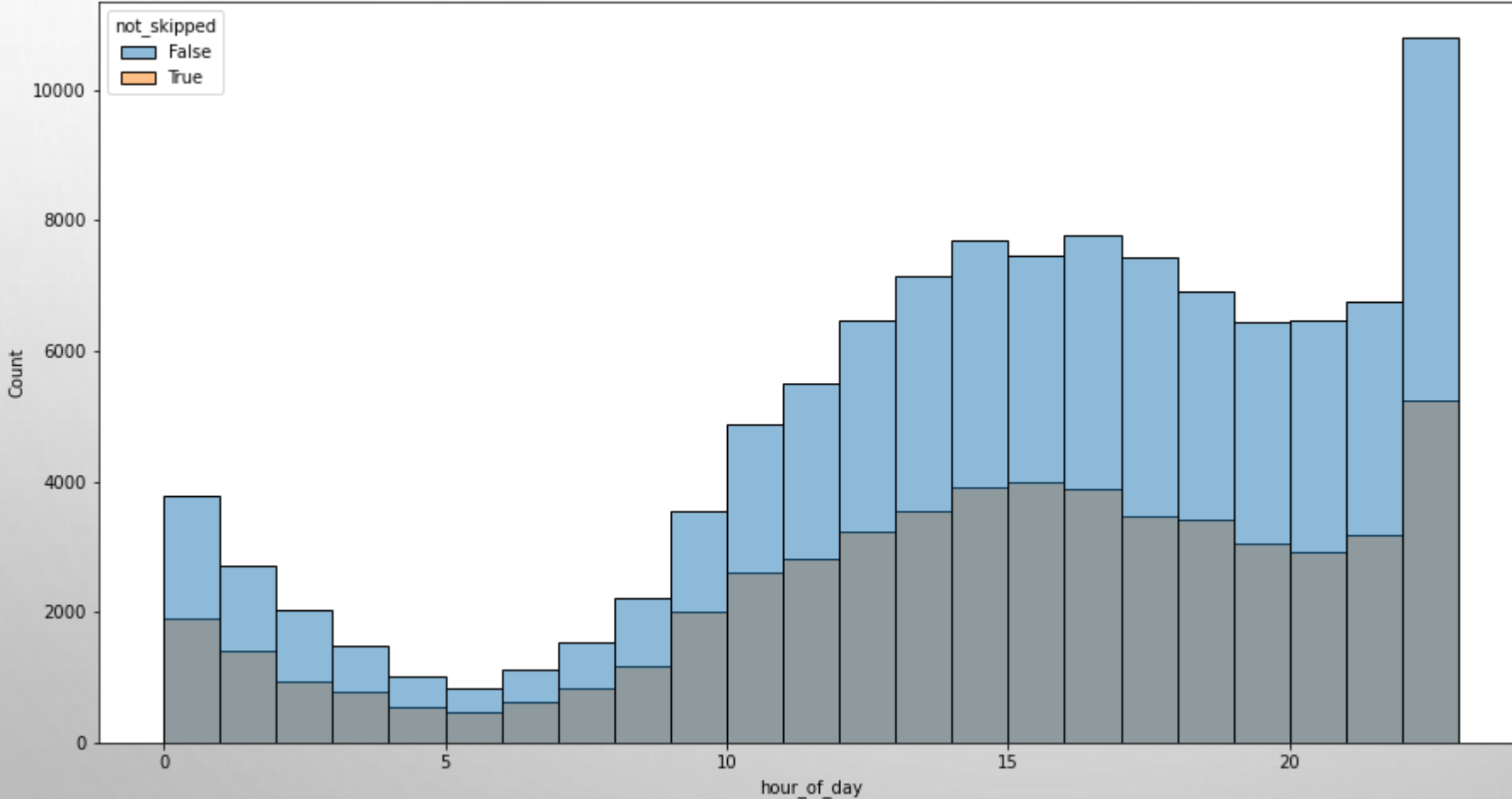# CAN LOUDNESS OF SONG IMPACT THE SKIP?

Songs are being skipped in phase 2 and 3 more. But the count of skips is normally distributed over loudness

CAN THE POPULARITY OF SONG IMPACT SKIP?

Seems like more the song's popularity More are they being skipped. But on the other hand, in each bin around 50% of the songs are being skipped.

CAN THE HOUR OF DAY IMPACT SKIP?

Time slot of 22-midnight registers the maximum song requests. But overall proportion of skips remains around that 50% mark.

PCA

- UNDERSTANDING THE IMPACT OF 70 FEATURES IS VERY DIFFICULT.

- TRANSFORMED THOSE 70 FEATURES INTO 4 PRINCIPLE COMPONENTS.

- THE VARIABILITY EXPLAINED BY THOSE COMPONENTS IS AROUND $0.9838960 \approx 98\%$

	pc_1	pc_2	pc_3	pc_4
0	1.103427	-0.564980	0.106102	0.011347
1	0.841644	0.919889	-0.396460	-0.089065
2	0.921508	-0.451135	-0.429430	-0.009270
3	0.928626	-0.548147	-0.432088	-0.023350
4	0.905070	-0.448440	-0.432234	-0.018499
5	-0.046544	0.772695	0.248712	-0.183551
6	-0.954521	-0.619079	0.046760	-0.262695
7	-0.965195	-0.670428	0.048839	-0.272748
8	-1.049345	0.786274	0.074048	-0.359910
9	-1.039047	0.761823	0.071510	-0.356322

Model SCORECARD

	Explanibility	Execution Time	Practical to implement	Accuracy
Logistic Regression	●●●●●	●●○○○	●●●●○	0.8049
KNN	●●●●○	●●●●○	●●●○○	0.8054
Random Forest	●●●○○	●●●○○	●●●●○	0.9274
SVM	●●○○○	●●●●●	●●○○○	0.6380

Thank You