Predicting Spotify Track Skipping

By Heather Knudson

About the Data

- Originally released for Spotify's 2018 Sequential Skip Prediction Challenge
- User behavior/context, *not* track features
- 130M observations ⇨ 2.9M subsample

Project Objective

To predict track skipping solely from user behavior/characteristics, rather than track features

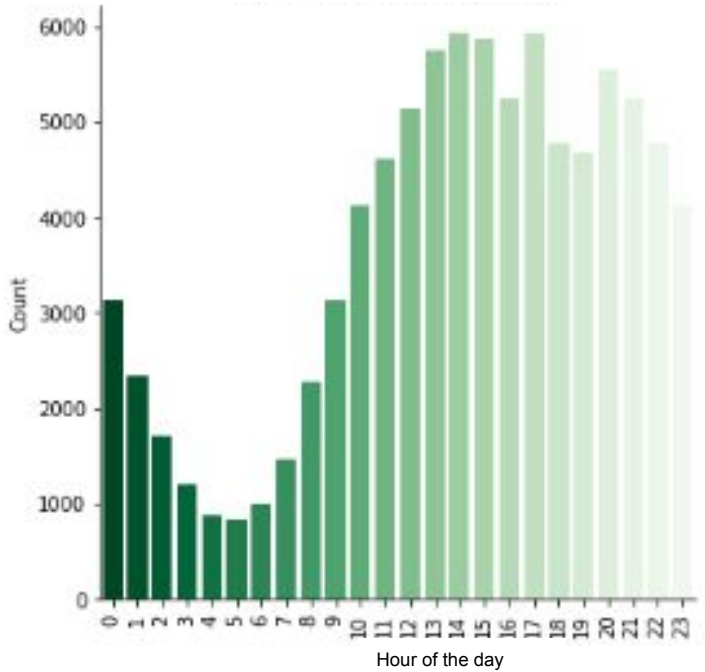
Data Cleaning

1. All categorical variables made numeric
2. Re-coded numerous variables to reverse negation
3. One-hot encoded all variables before modelling
4. Filtered data to include only listening sessions with 20 tracks
5. Filtered again to include only observations that were either the 1st, 10th, or 20th track in a listening session

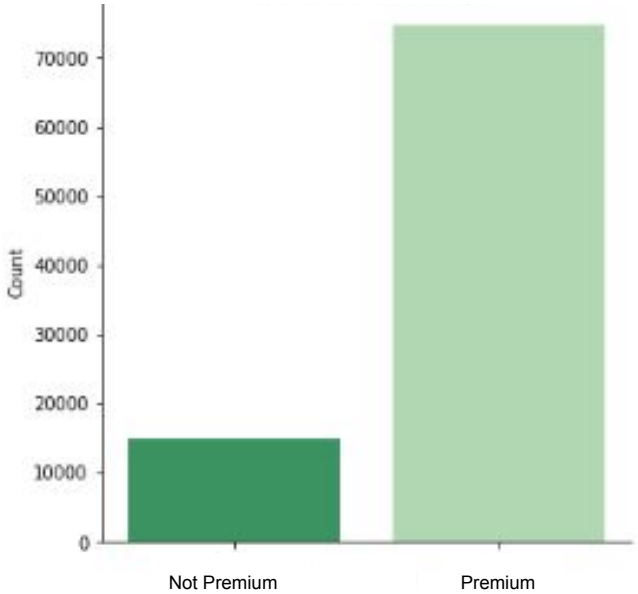
3 datasets ● 89,672 obs in each ● 23 columns

Descriptive Statistics

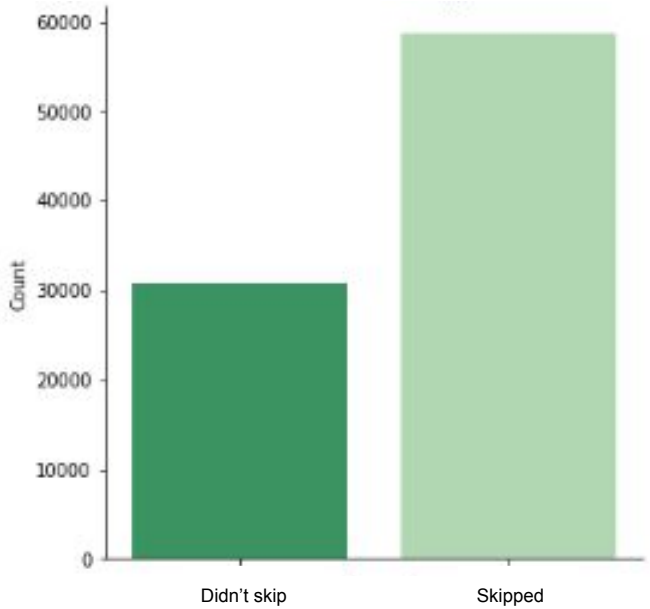
Most listening sessions were in the afternoon and night



Most Spotify users are premium members



Most users skipped their current track

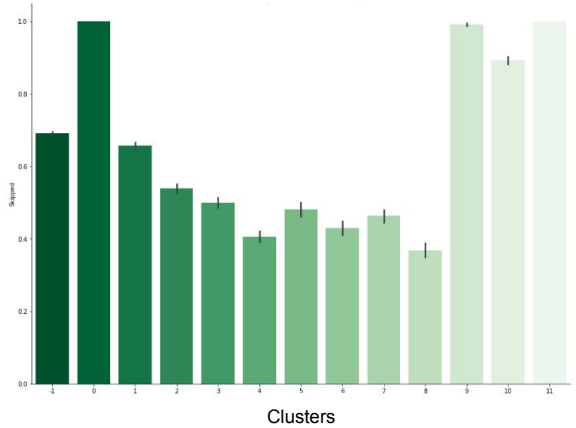


Clustering

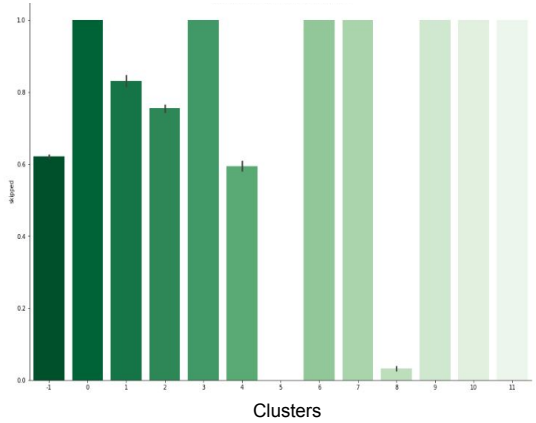
- HDBSCAN
- 12-cluster solution for all 3 tracks

The clusters help explain the variation in skipping across tracks

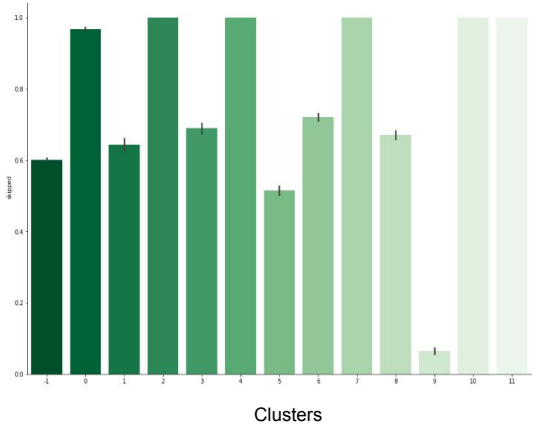
Track 1



Track 10



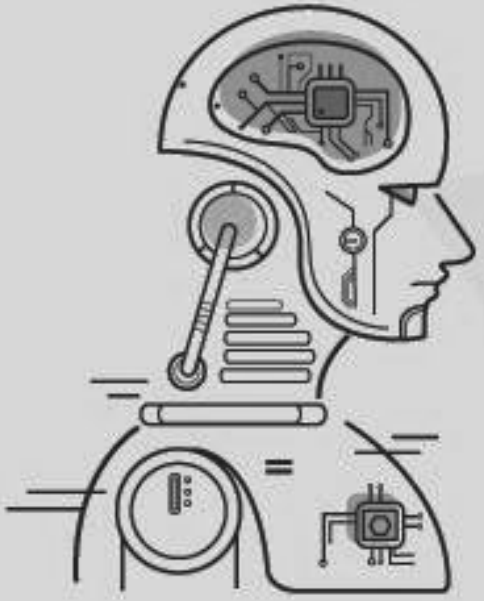
Track 20

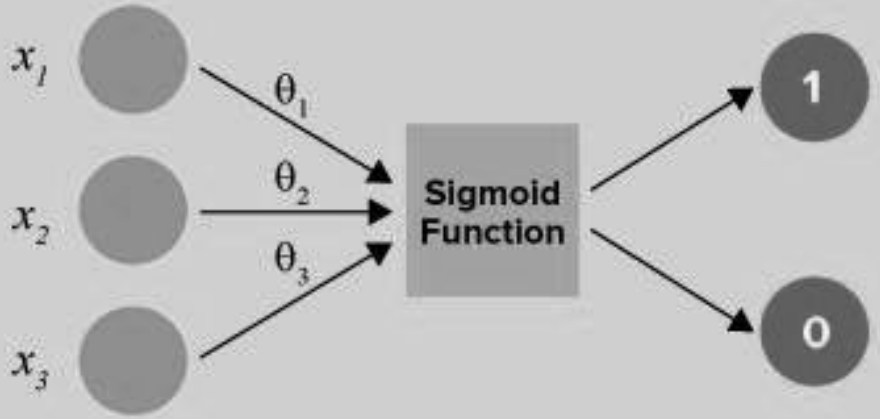


Methods

- **Feature selection**
- **Model optimization**
- **Model evaluation**
- **Model interpretation**

Logistic Regression



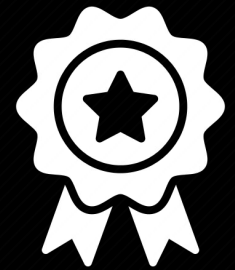


Best Scores from Logistic Regression

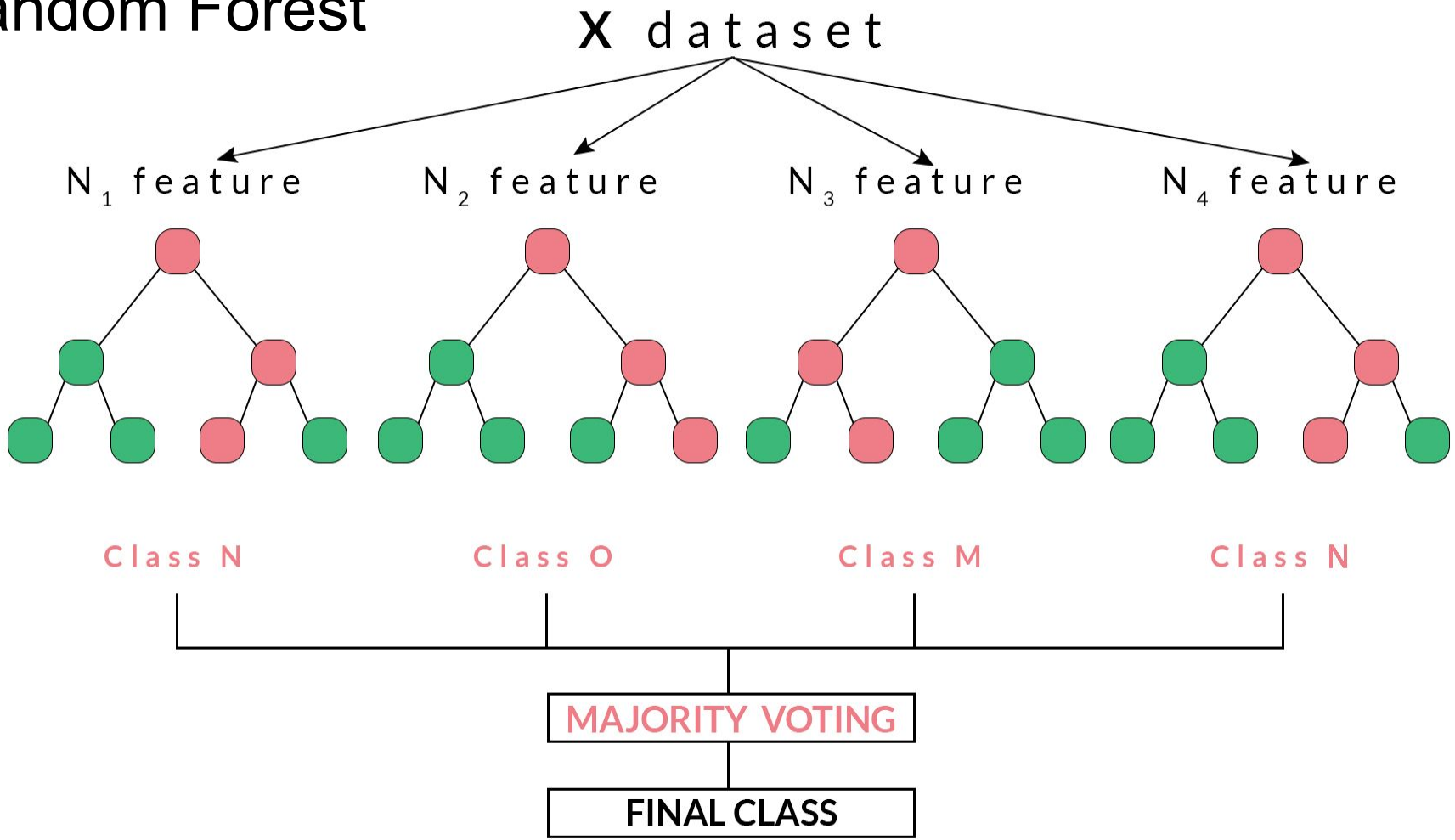
Track 1: 72% AUC; 17% & 79% F1; 66% accuracy

Track 10: 86% AUC; 68% & 89% F1; 84% accuracy

Track 20: 85% AUC; 70% & 88% F1; 82% accuracy



Random Forest

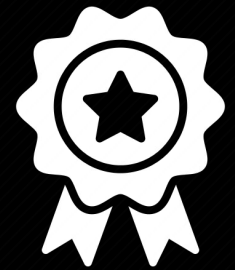


Best Scores from Random Forest

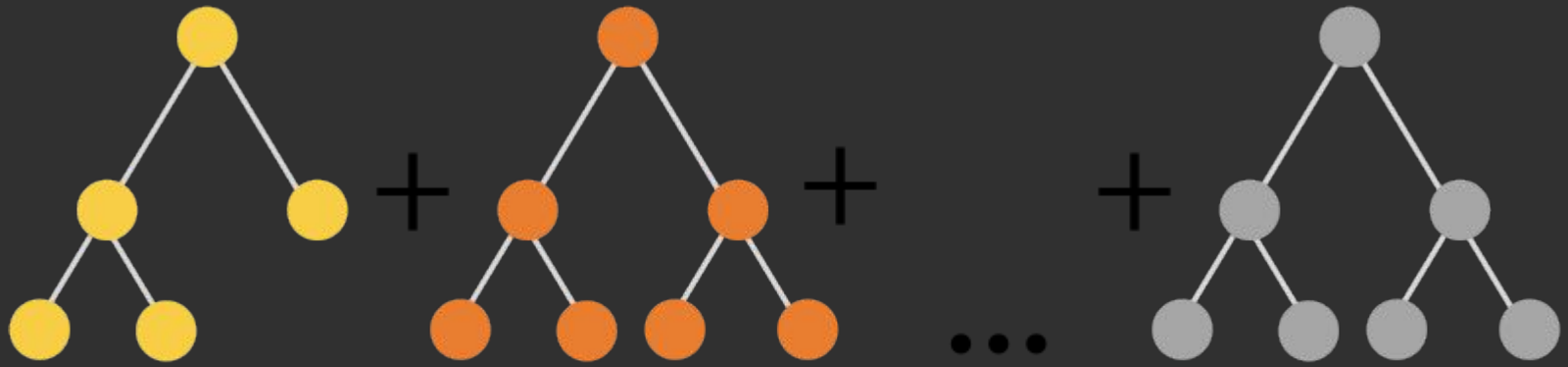
Track 1: 80% AUC; 58% & 79% F1; 72% accuracy

Track 10: 95% AUC; 81% & 93% F1; 89% accuracy

Track 20: 93% AUC; 79% & 91% F1; 87% accuracy



XGBoost



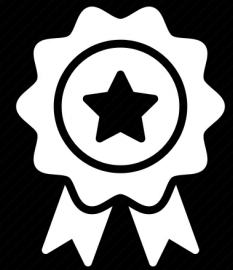
Best Scores from XGBoost

Final Model

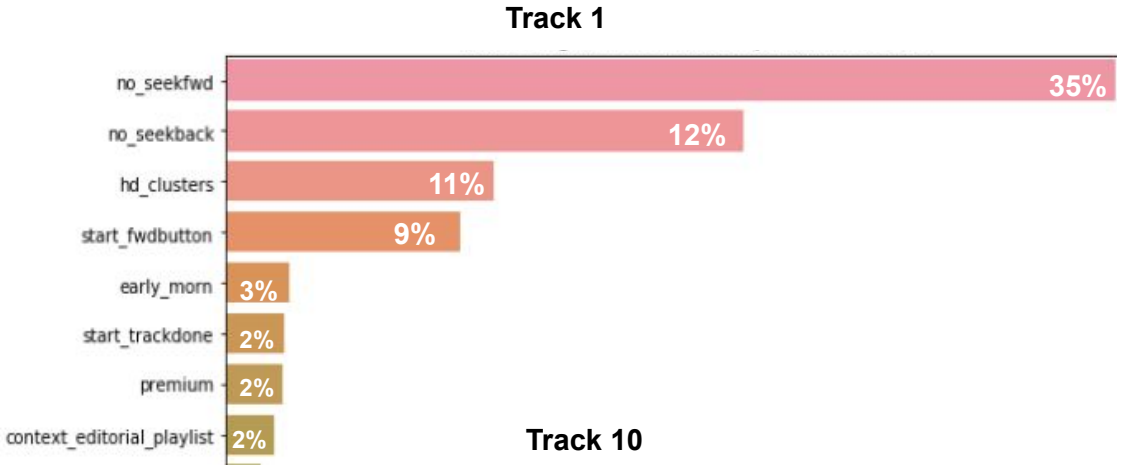
Track 1: 80% AUC; 66% & 73% F1; 70% accuracy

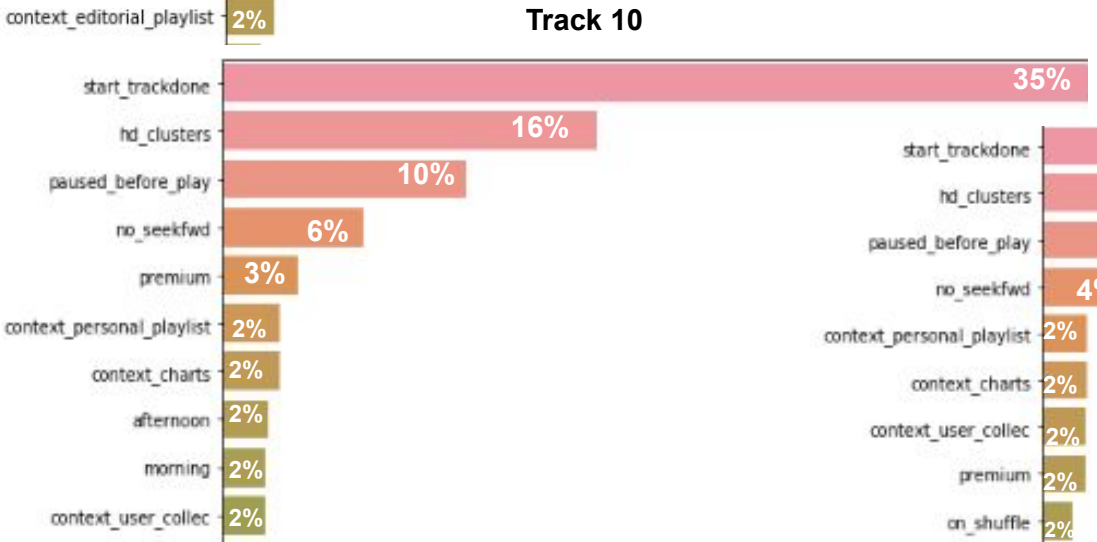
Track 10: 95% AUC; 81% & 93% F1; 89% accuracy

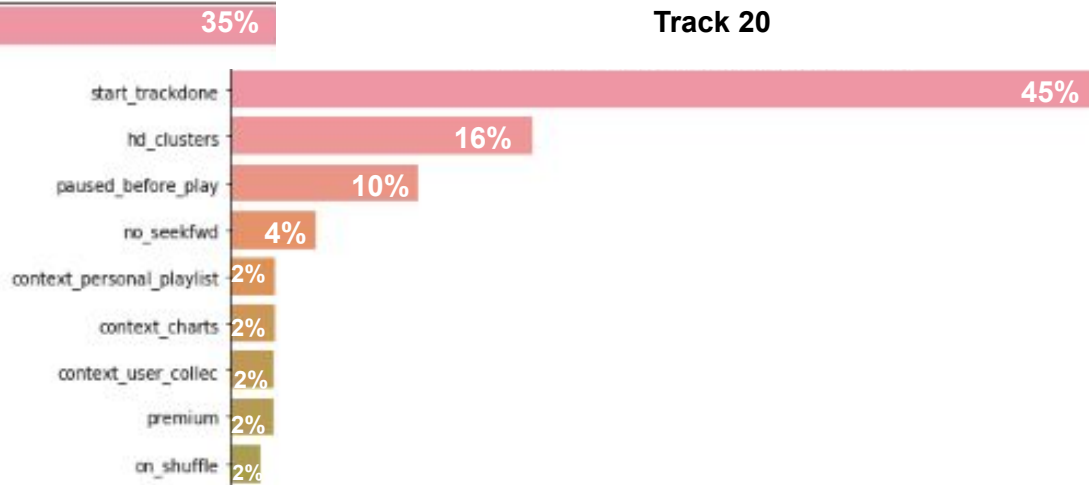
Track 20: 93% AUC; 79% & 91% F1; 87% accuracy



Feature Importance







Odds Ratios

Hd_clusters

The clusters that resulted from each track's HDBSCAN solution

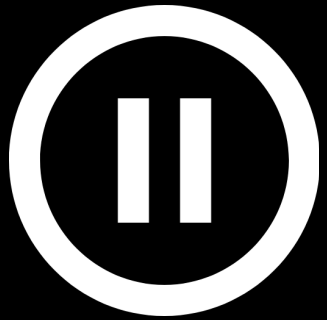
Track	Odds Ratio	Likelihood of Skipping
Track 1	1.0	Equally as likely
Track 10	1.2	20% more
Track 20	1.1	10% more

Odds Ratios

Paused_before_play

The current track was paused before it was played

Track	Odds Ratio	Likelihood of Skipping
Track 1	N/A	N/A
Track 10	17.0	17 times more
Track 20	12.7	12.7 times more



Odds Ratios

No_seekfwd

The user did not seek forward within their current track

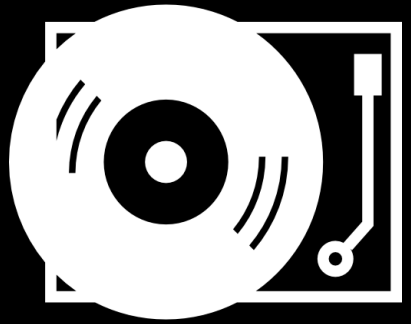
Track	Odds Ratio	Likelihood of Skipping
Track 1	.18	82% less likely
Track 10	.09	91% less likely
Track 20	.07	93% less likely

Odds Ratios

Start_trackdone

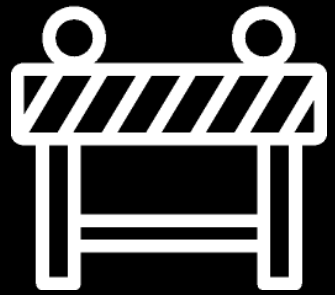
The current track started because the previous track ended

Track	Odds Ratio	Likelihood of Skipping
Track 1	.803	19.7% less
Track 10	.040	96% less
Track 20	.045	95.5% less



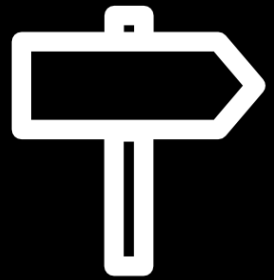
Limitations

- Track 1 difficult to optimize
- Categorical data
- Relatively low amount of computing power



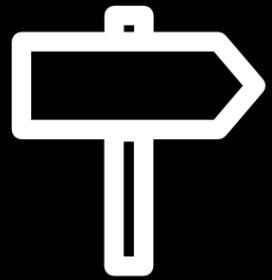
Conclusions

- The majority of Spotify users *did* skip their current track
- Pausing before playing was highly influential to skipping, as was fast-forwarding within the current track
- Users who listened to a personal playlist were much more likely to skip



Implications for Spotify

- Spotify is motivated to prevent skipping to improve song recommendations
- This dataset, however, looked at reasons other than track information to explain why skipping occurs
- This could be useful in determining confounding variables *other than track features* that affect skipping



Future Research Possibilities

- Another project that uses just the original variables rather than one-hot encoding
- Use this information on user behaviors in further investigation of track features to determine if these user-related variables could be confounding variables