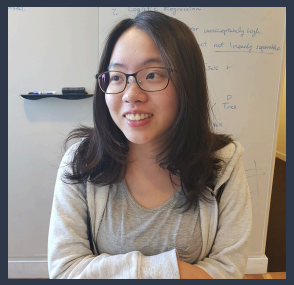
Practical Data Science Final Project

(2018 Spring)

Prediction on whether
a Hit Song* Has the Potential to Become *a Super Hit



Amy Huang

Wei-Kung Wang

Chia-Hua Lee

Data Collection

billboard CHARTS NEWS VIDEO PHOTOS BUSINESS Subscribe

Festivals Hot 100 Billboard 200 Latin Podcasts Pop R&B/Hip-Hop Chart Beat Artists

THE HOT 100

GET NOTIFIED! Music moves fast. Don't miss out. Get Billboard chart news first. TRY IT NOW

How it works

Search the Chart Archives. Try entering a birthday or anniversary.

THE WEEK OF MAY 12, 2018 ARCHIVE SEARCH MM/DD/YYYY

1 Last Week: 1

Nice For What
Drake

2 Last Week: 5

Psycho
Post Malone Featuring Ty Dolla $ign

3 Last Week: 2

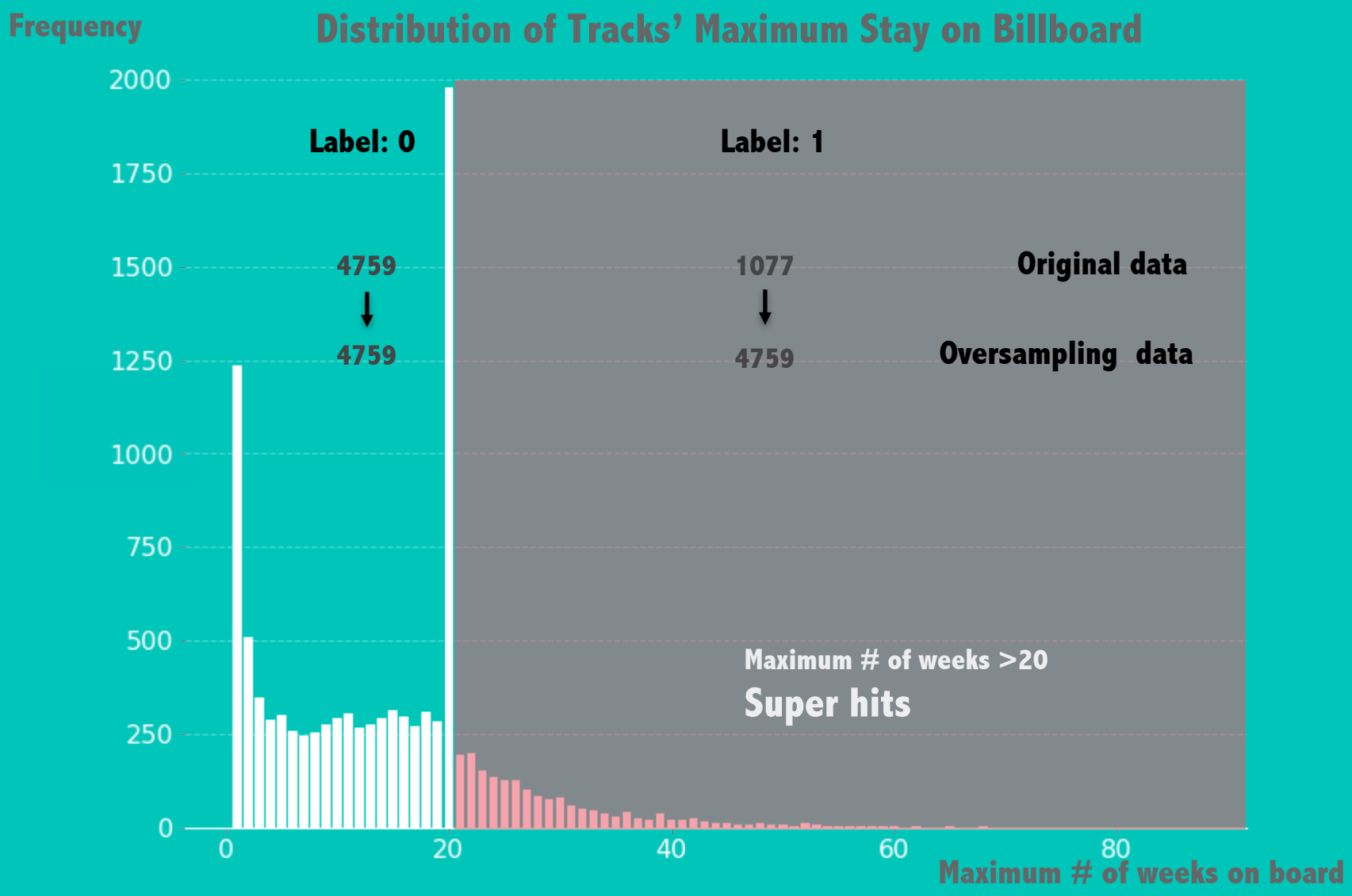
God's Plan
Drake

1990-2018

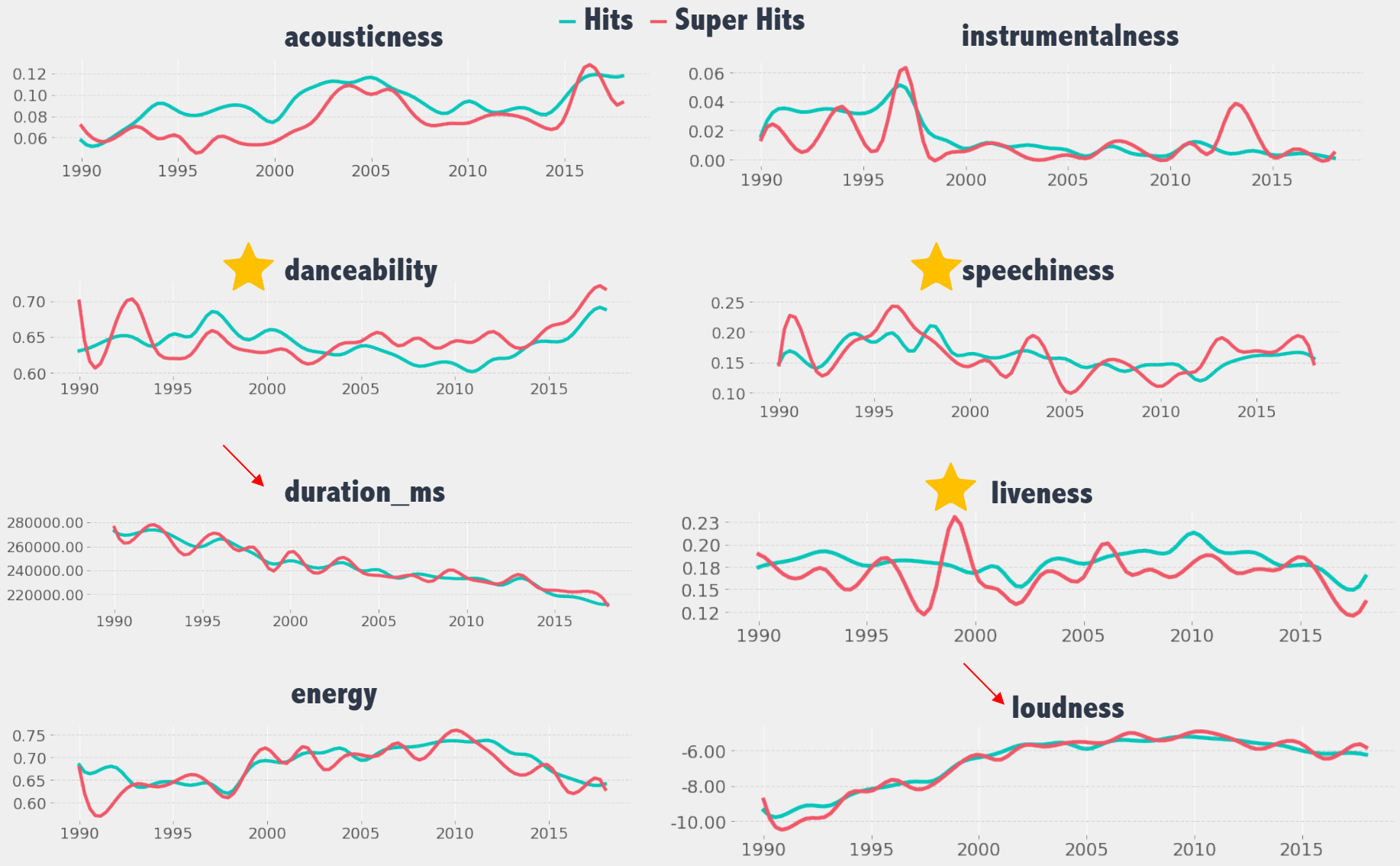
Spotify

Get Audio Features & Metadata

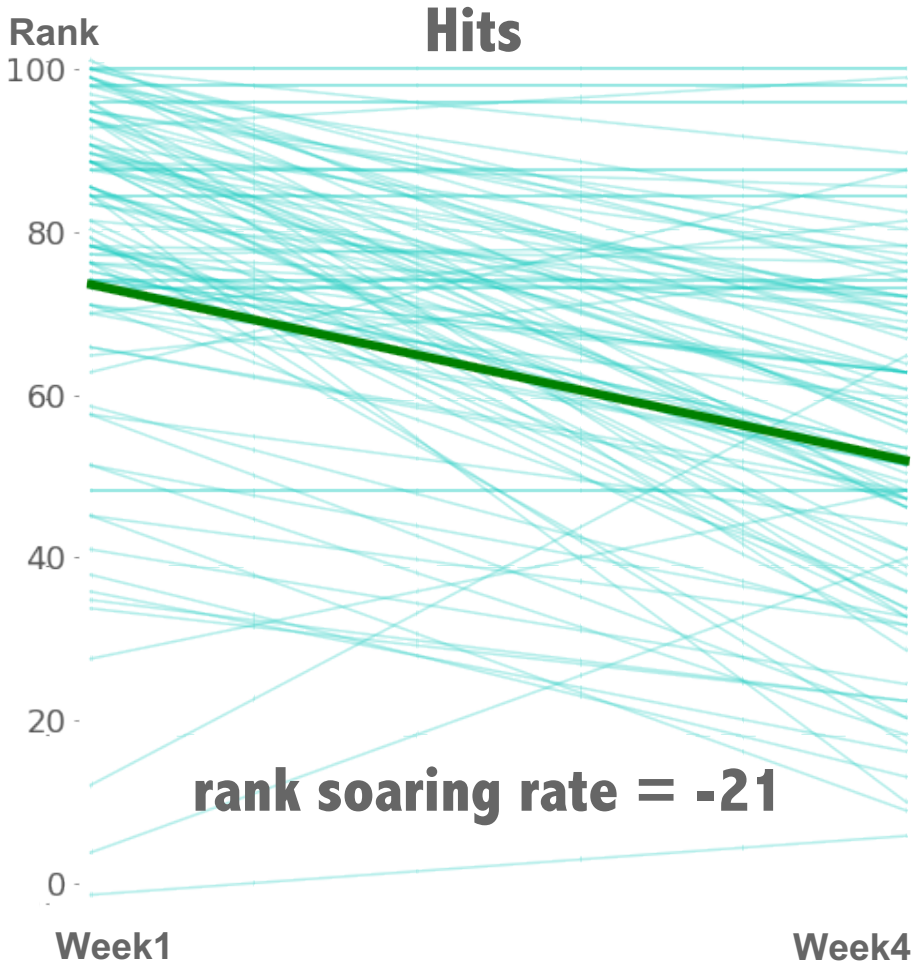
Defining the Super Hits and Preprocessing Data

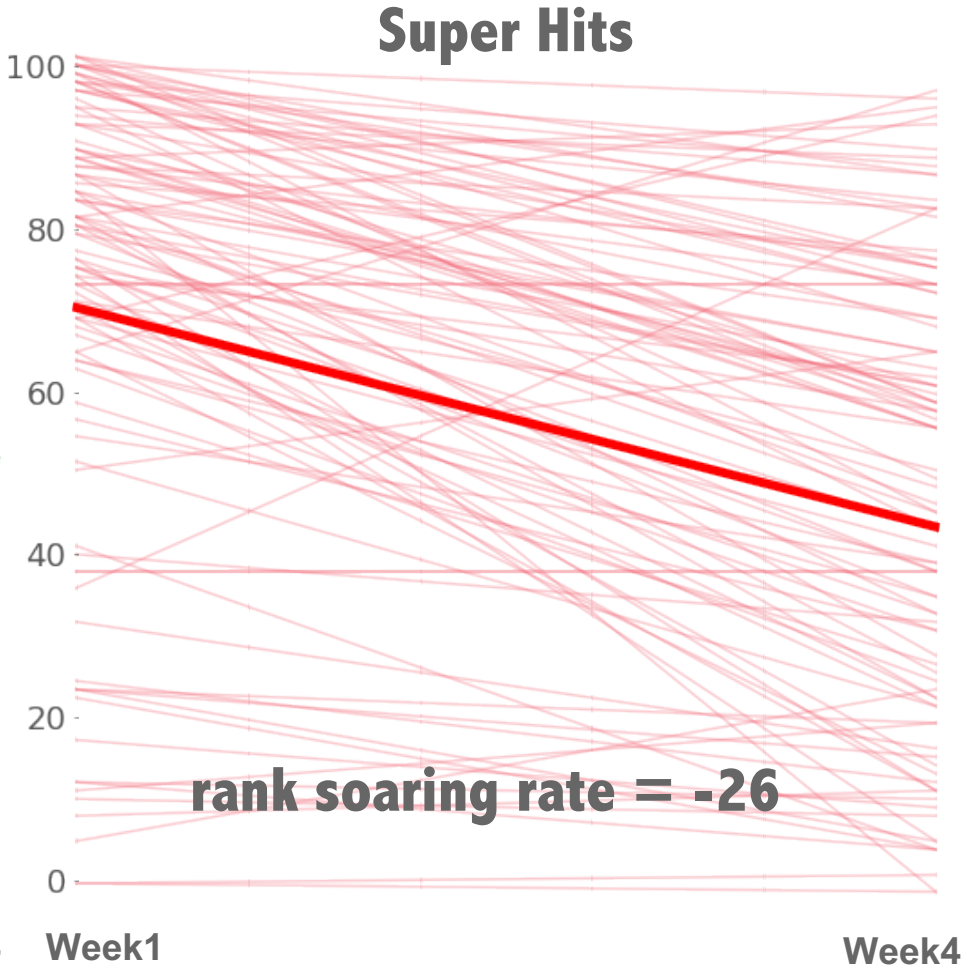


Audio Features of Hits vs Super Hits over Years



Engineering a New Feature!

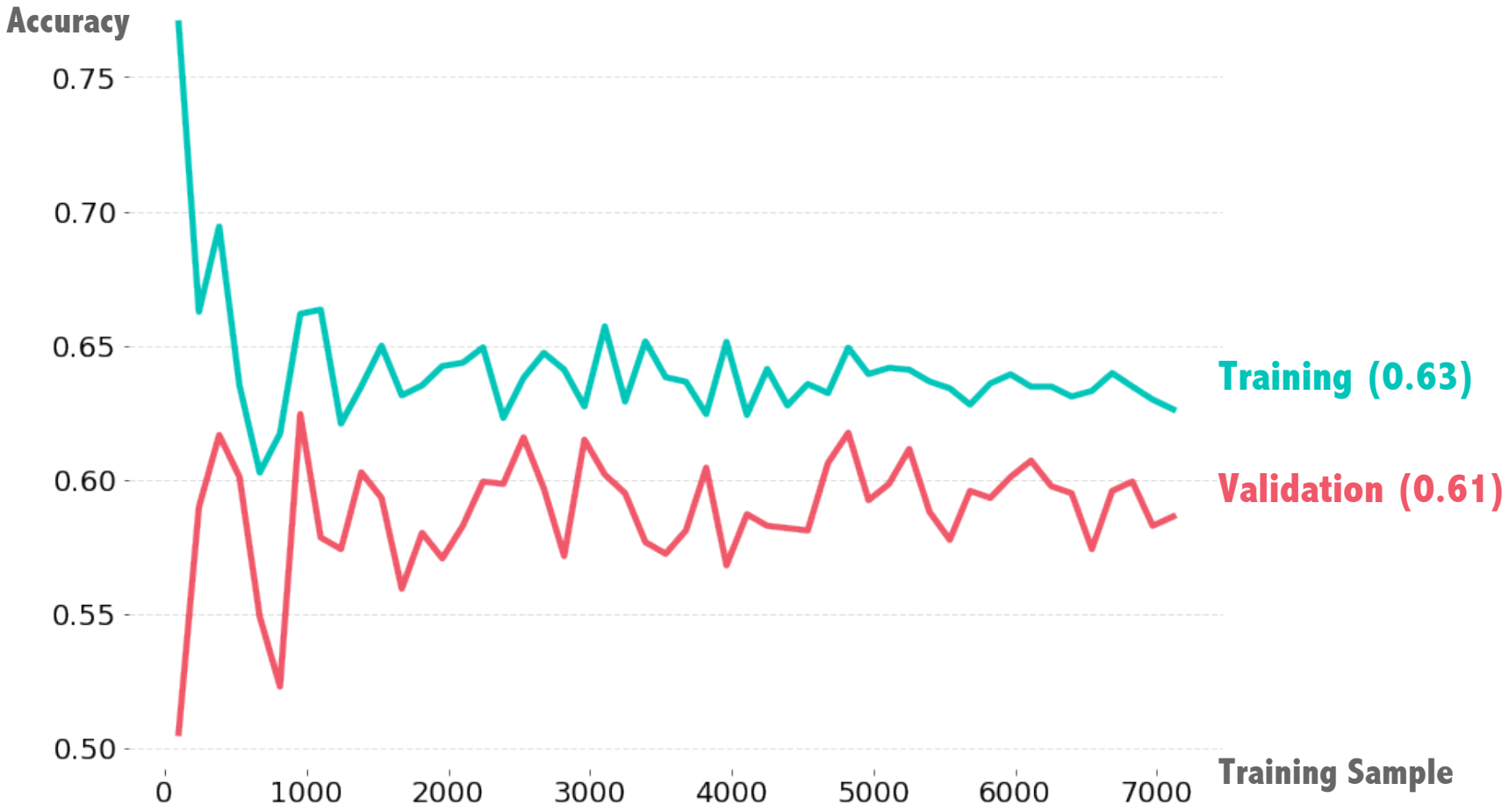




Feature Engineering :

Monthly Rank Soaring Rate = Rank of Week4 – Rank of Week1

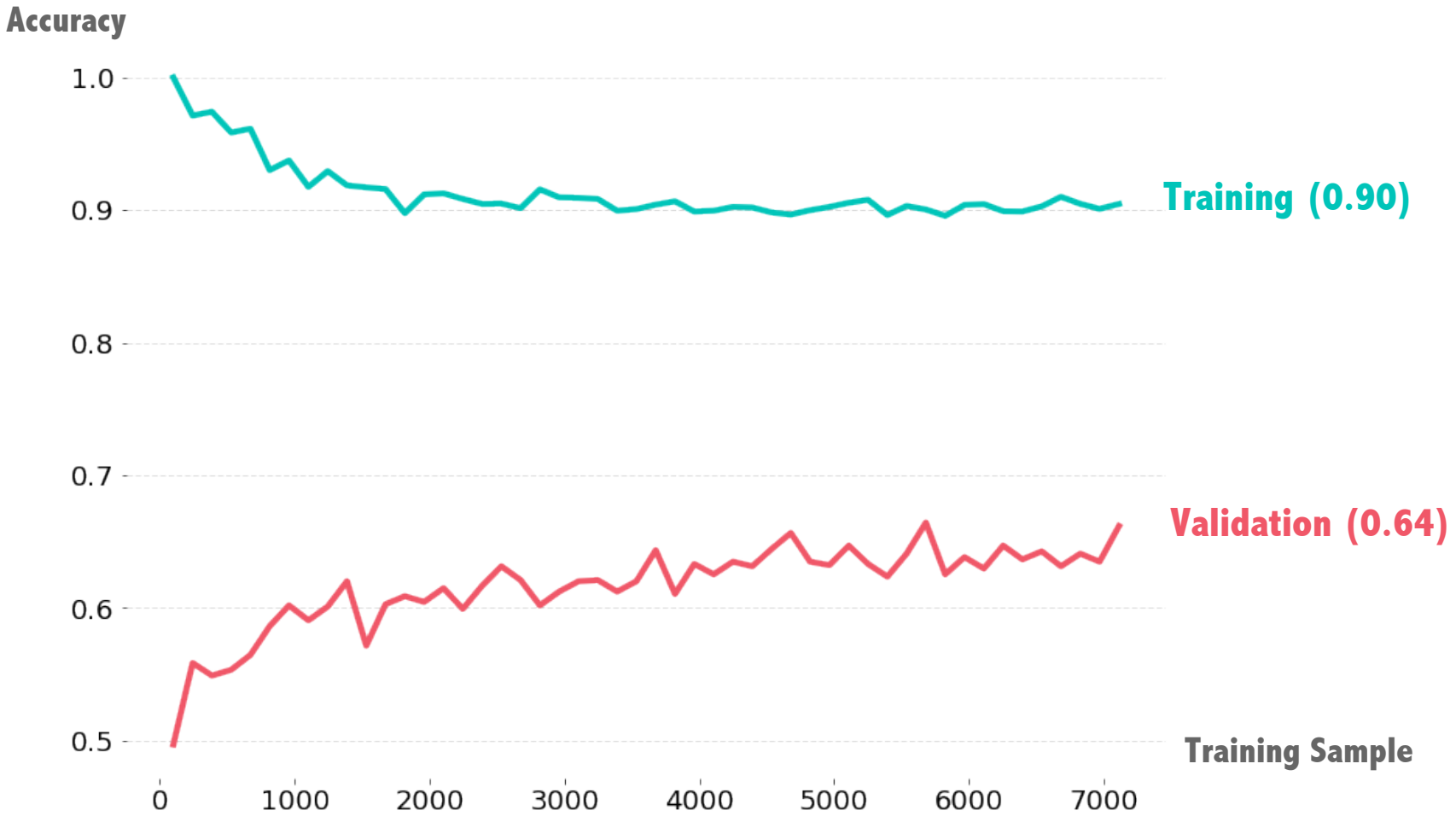
Model Building: Logistic Regression (Linear Classifier)



☹ **Low training accuracy on Logisic Regrsssion indicates a non-linearly separable dataset**

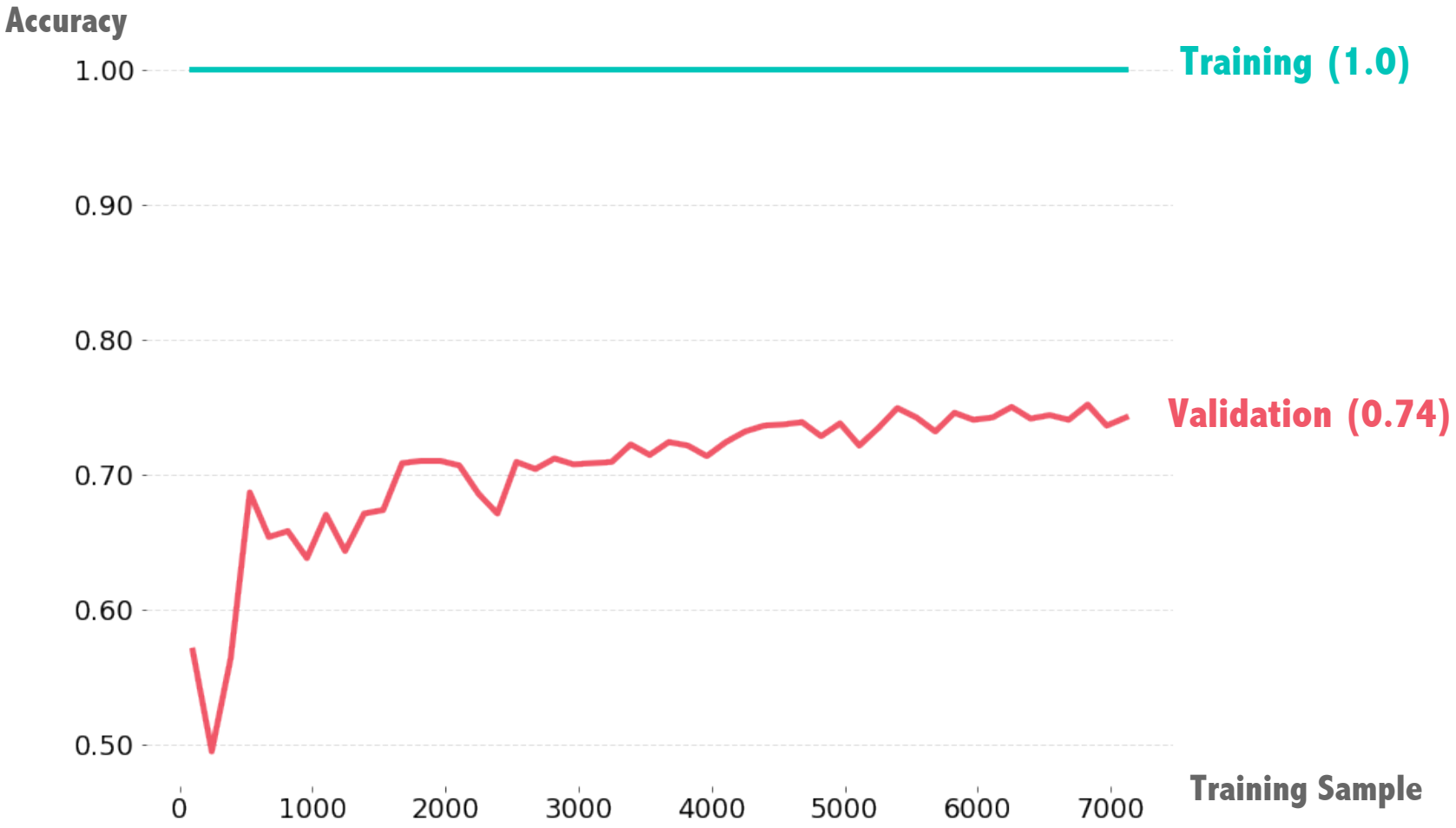
→ **Try non-linear classifiers**

Model Building: RBF Kernel SVM (Nonlinear Classifier)



High training accuracy with low validation accuracy show high variance in model

Model Building: Random Forest



Rising validation accuracy with sample size → A hope of better result with more samples

Top predictors from Random Forest matches what we saw previously in EDA

Model Evaluation & Conclusion

	Accuracy	Precision	Recall	F1 Score
Baseline*	0.64	0.17	0.15	0.16
Logistic Regression	0.61	0.30	0.54	0.39
Kernel SVM	0.65	0.31	0.46	0.37
★ Random Forest	0.75	0.44	0.30	0.36

Baseline: Classify a song as Super Hit if the rank is rising in the first month.

There is hope if we have more data!