Spotify music ranking prediction

Ville Saarinen
Sirong Huang

DATA & GOAL

Target: **total streams** on Spotify

Danceability

Tempo, rhythm stability, beat strength, regularity

Energy

Death metal vs Bach

Acousticness

Valence

Happy, cheerful, euphoric
Sad, depressed, angry

Tempo

Speed, beat duration

Loudness

Exploratory data analysis

Figure 1. Feature pairwise correlation plot

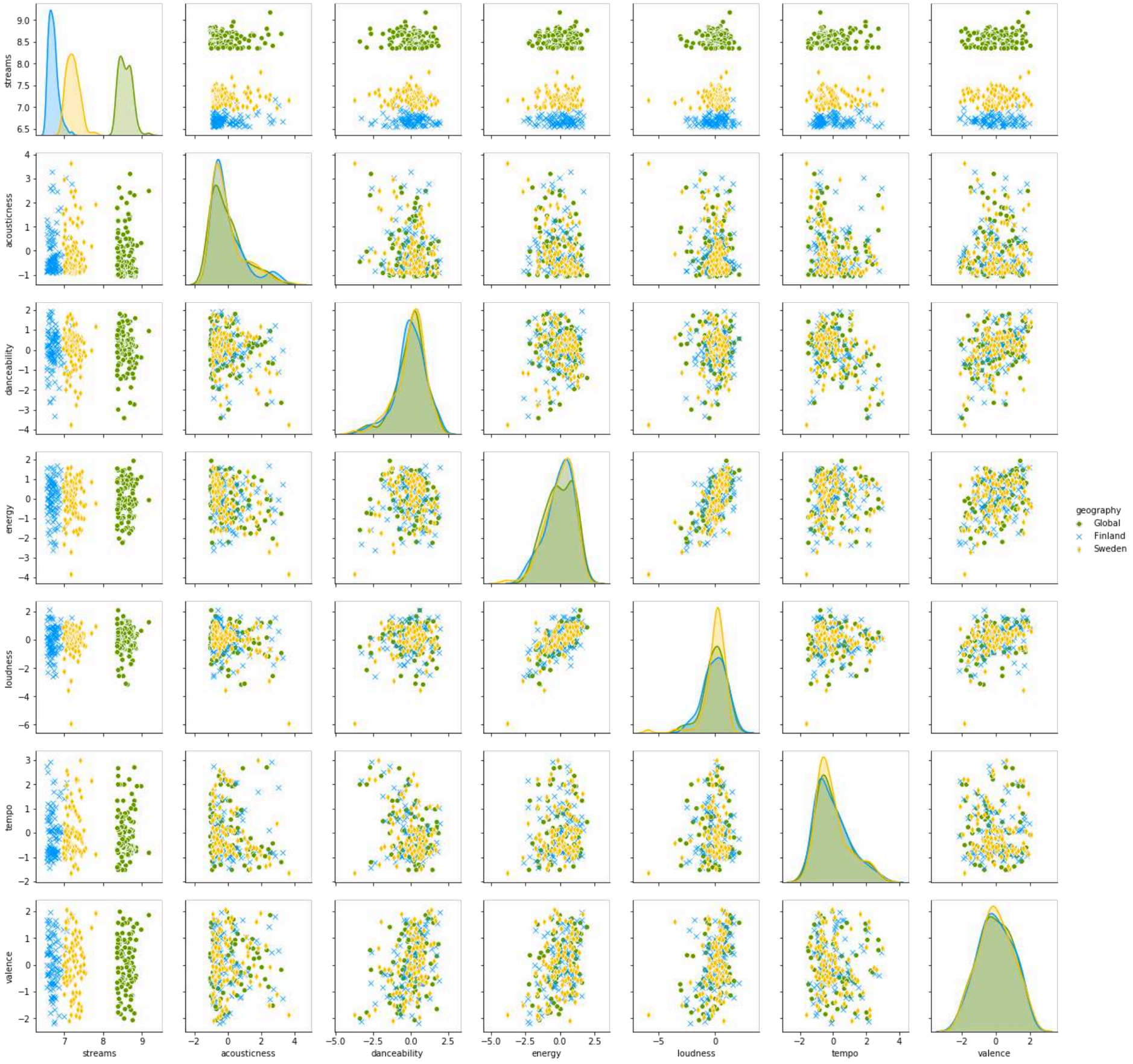


Figure 2. Feature pairwise scatterplot

Model & prior choice

- Linear regression with 3 predictors

$$Y \sim N(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_d X_d,\ \sigma)$$

- Non-linear regression with interaction terms with 5 predictors

$$Y \sim N(\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_2 X_4 + \beta_7 X_3 X_5 + \beta_8 X_1 X_5 + \beta_9 X_2 X_5,\ \sigma)$$

- Uninformative and weakly informative priors

$$\alpha \sim N(0, 100\sigma_0^2), \text{ and } \beta_j \sim N(0, 100\sigma_0^2).$$

Model diagnostics

Model 1: Linear regression with 3 predictors

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
a	8.774083	0.000293	0.018475	8.737628	8.762250	8.773631	8.786050	8.811840	3977.814460	0.999426
b[1]	0.024649	0.000237	0.015494	-0.006393	0.014677	0.024506	0.035100	0.054683	4291.545842	0.999662
b[2]	0.026977	0.000285	0.018368	-0.010112	0.014825	0.027040	0.039314	0.062038	4150.062894	0.999482
b[3]	0.045904	0.000391	0.024032	-0.002407	0.030046	0.046332	0.061807	0.092253	3784.439618	0.999822
sigma	0.097116	0.000258	0.014488	0.073484	0.086797	0.095401	0.105654	0.130574	3163.271457	1.000832

Model	loo	loo_se	p_loo	warning	div	treedepth	energy
Model 1: 3 Predictors, Uninformative	330.12607	1193.251565	1216.707605	1	False	True	True

Figure 3. Predictive performance of Model 1

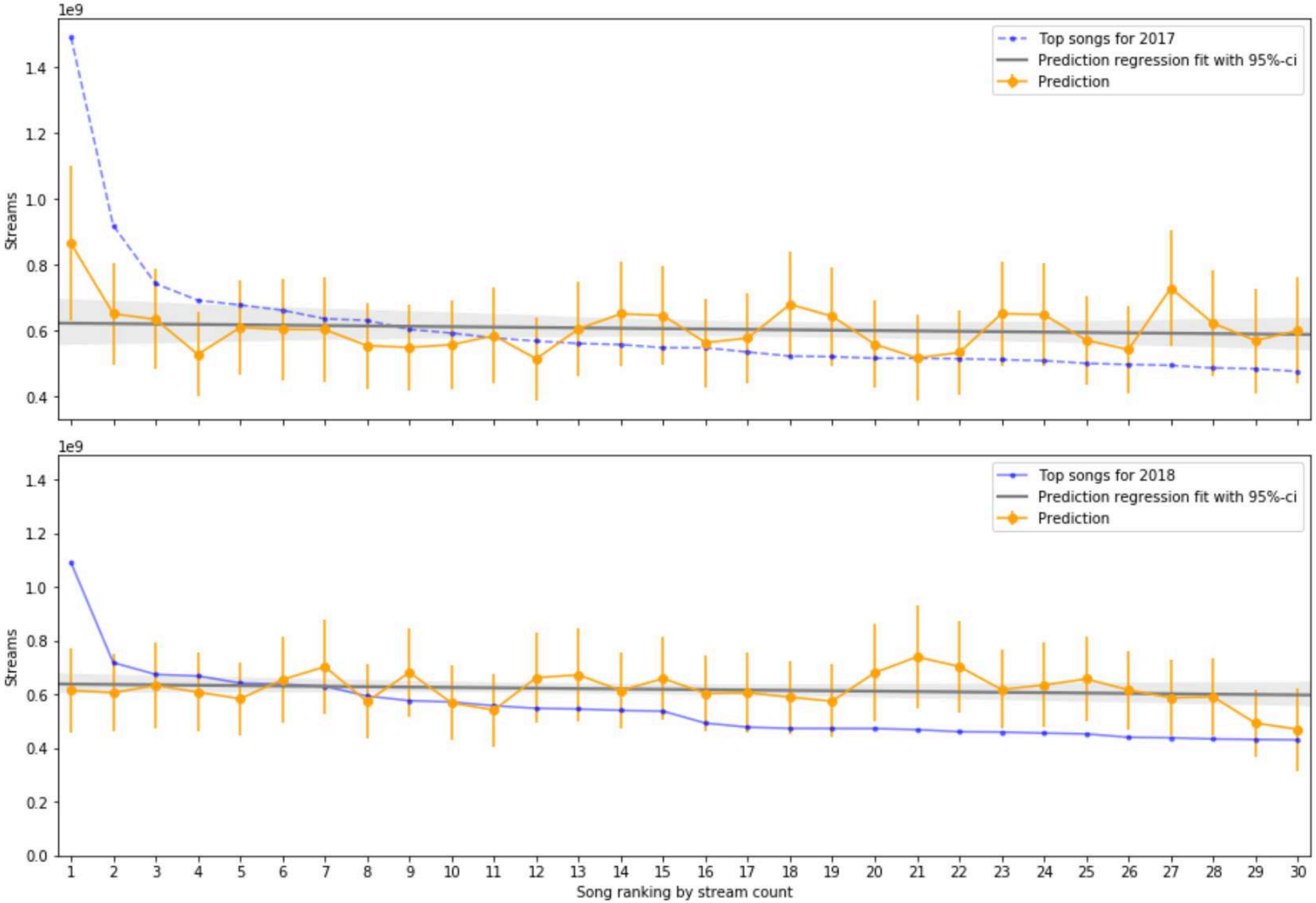
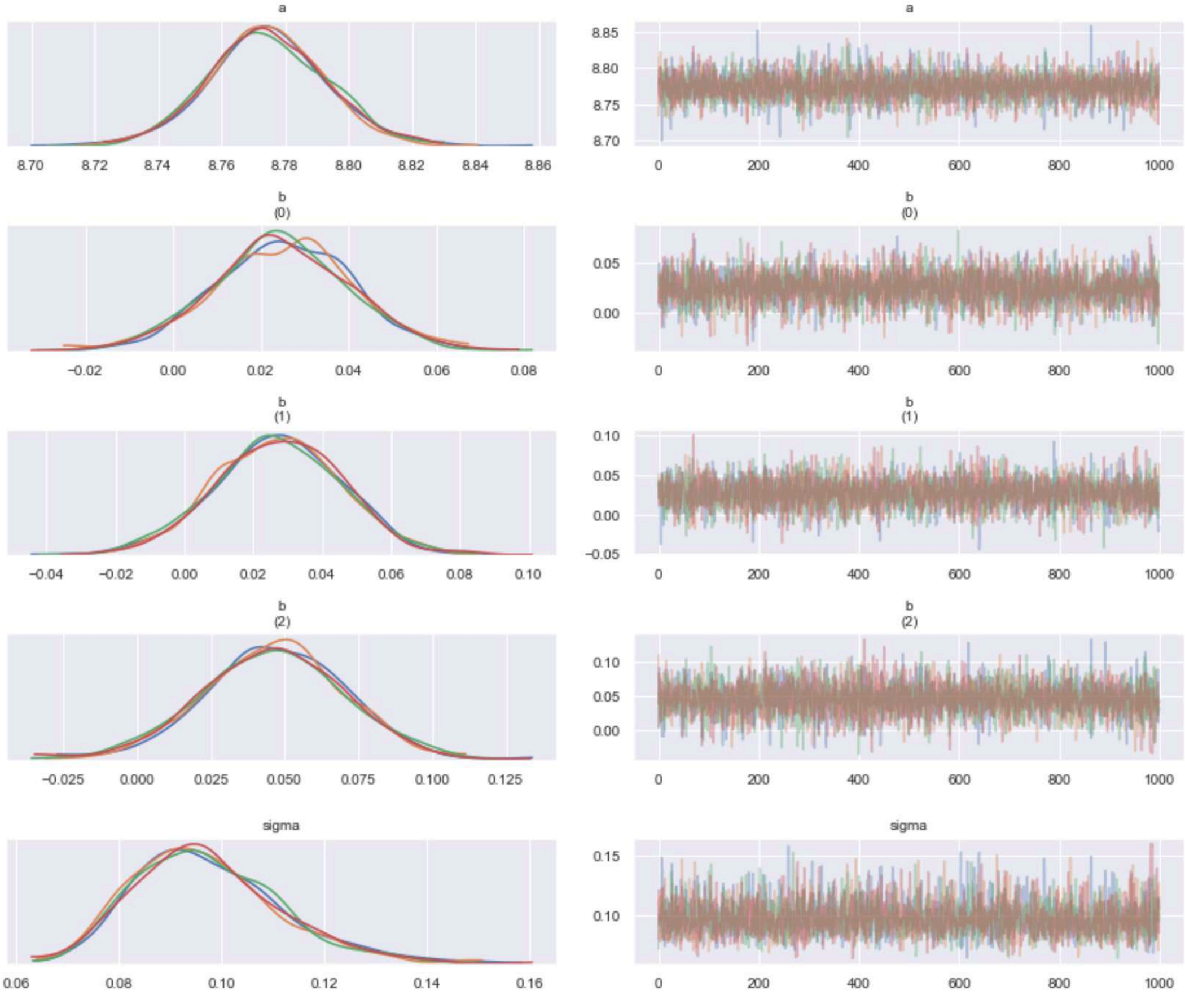


Figure 4. Trace plots of model 1



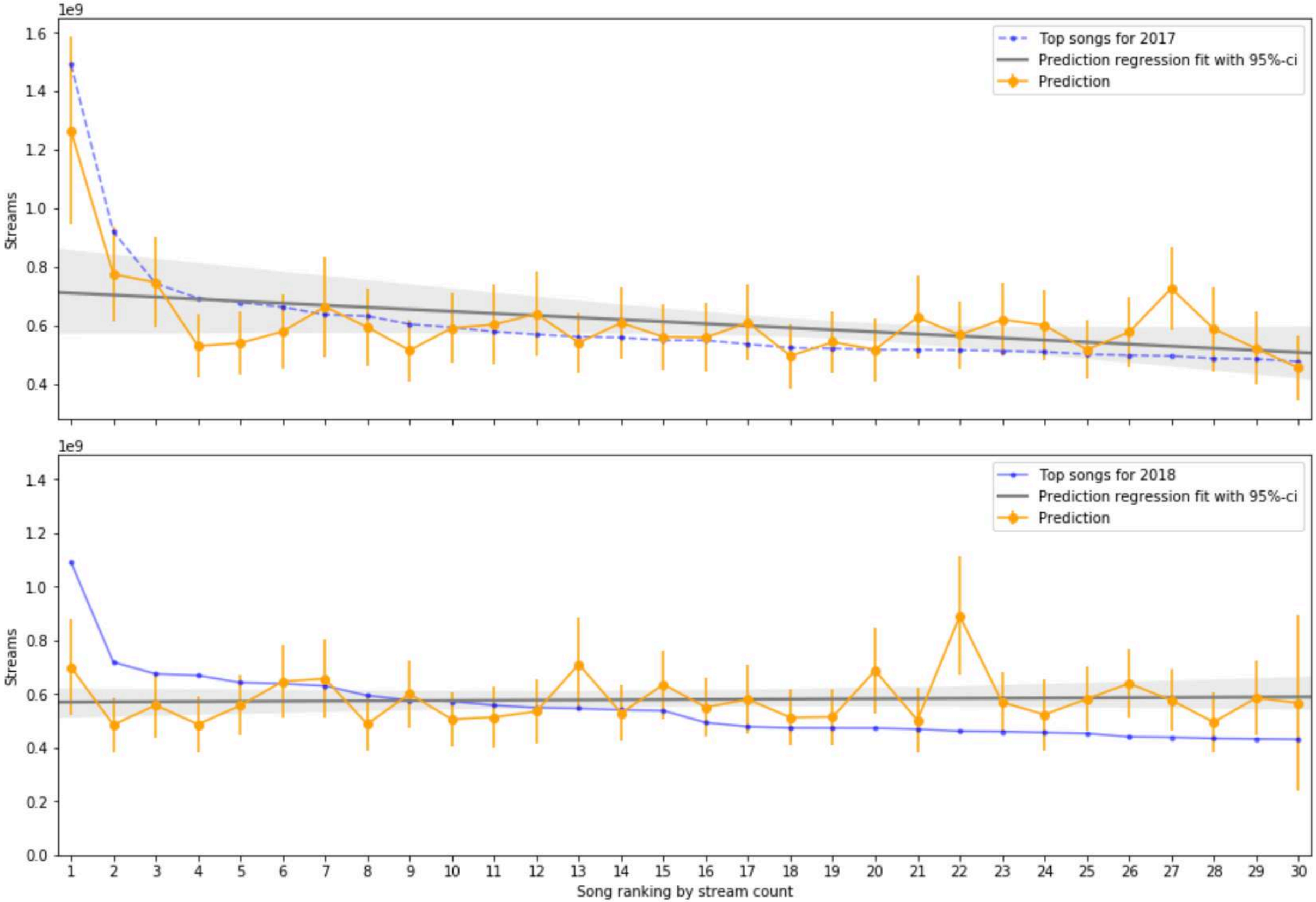
Model diagnostics

Model 2: Linear regression with 3 predictors

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
a	8.728424	0.000389	0.020158	8.688706	8.715223	8.728735	8.741788	8.767825	2686.726787	1.002356
b[1]	0.003220	0.000322	0.016158	-0.028904	-0.007326	0.003327	0.013890	0.034465	2511.786295	1.001377
b[2]	-0.002634	0.000388	0.019684	-0.041825	-0.015390	-0.002223	0.010088	0.035605	2572.578450	1.001855
b[3]	0.023101	0.000529	0.027000	-0.030475	0.005858	0.022815	0.040605	0.076748	2605.178369	1.000530
b[4]	0.027016	0.000383	0.019940	-0.013155	0.014176	0.027138	0.039676	0.065983	2711.904930	1.000195
b[5]	0.029483	0.000419	0.020094	-0.009965	0.016512	0.029381	0.042268	0.069783	2296.859189	1.000317

	loo	loo_se	p_loo	warning	div	treedepth	energy
Model 2: 5 Predictors with 2 interaction terms	3072.952117	1970.499454	2411.396402	1	True	True	True

Figure 5. Predictive performance of Model 2



Model diagnostics

Model 2: Linear regression with 3 predictors

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
a	8.728424	0.000389	0.020158	8.688706	8.715223	8.728735	8.741788	8.767825	2686.726787	1.002356
b[1]	0.003220	0.000322	0.016158	-0.028904	-0.007326	0.003327	0.013890	0.034465	2511.786295	1.001377
b[2]	-0.002634	0.000388	0.019684	-0.041825	-0.015390	-0.002223	0.010088	0.035605	2572.578450	1.001855
b[3]	0.023101	0.000529	0.027000	-0.030475	0.005858	0.022815	0.040605	0.076748	2605.178369	1.000530
b[4]	0.027016	0.000383	0.019940	-0.013155	0.014176	0.027138	0.039676	0.065983	2711.904930	1.000195
b[5]	0.029483	0.000419	0.020094	-0.009965	0.016512	0.029381	0.042268	0.069783	2296.859189	1.000317

	loo	loo_se	p_loo	warning	div	treedepth	energy
Model 2: 5 Predictors with 2 interaction terms	3072.952117	1970.499454	2411.396402	1	True	True	True

Model comparison

	div	energy	loo	loo_se	p_loo	treedepth	warning
Linear; 3 Predictors Uninformative	False	True	330.126070	1193.251565	1216.707605	True	1
Linear; 3 Predictors Weakly Informative, Normal	True	True	414.275619	1201.617354	1270.207832	True	1
Linear; 5 Predictors Weakly Informative, Normal	True	True	568.040652	1219.729017	1318.150582	True	1
Non-Linear; 5 Predictors Weakly Informative inter 1, Normal	True	True	1692.046913	1611.779674	1805.222313	True	1
Non-Linear; 5 Predictors Weakly Informative inter 2, Normal	True	True	3072.952117	1970.499454	2411.396402	True	1
Non-Linear; 5 Predictors Weakly Informative squared, Normal	True	True	2255.697280	1633.004880	2001.678876	True	1
Non-Linear; 5 Predictors Weakly Informative squared inter, Normal	True	True	3231.741395	1890.545766	2446.777666	True	1
Non-Linear; 5 Predictors Weakly Informative cubic, Normal	True	True	7528.847211	3188.444969	4436.048235	True	1

Conclusion

- Model does not generalize well on new data
- Overfitting, high standard error, unreliable
- Increase model complexity does yield better result
- Data lacks explanatory power towards the prediction target

Limitation

- Song ranking is a determined by complex factors
- Inherent limitations in data
  - not sufficient to predict ranking
  - high variation across rankings
- Model unreliability and incapability to capture the relationship
  - LOO scores

Future improvement

- Gather more relevant data
  - marketing activities, social media
  - artists, genre etc.
- Informative prior with domain expertise
- Hierarchical model

Spotify

Thank you!

Ville Saarinen
Sirong Huang