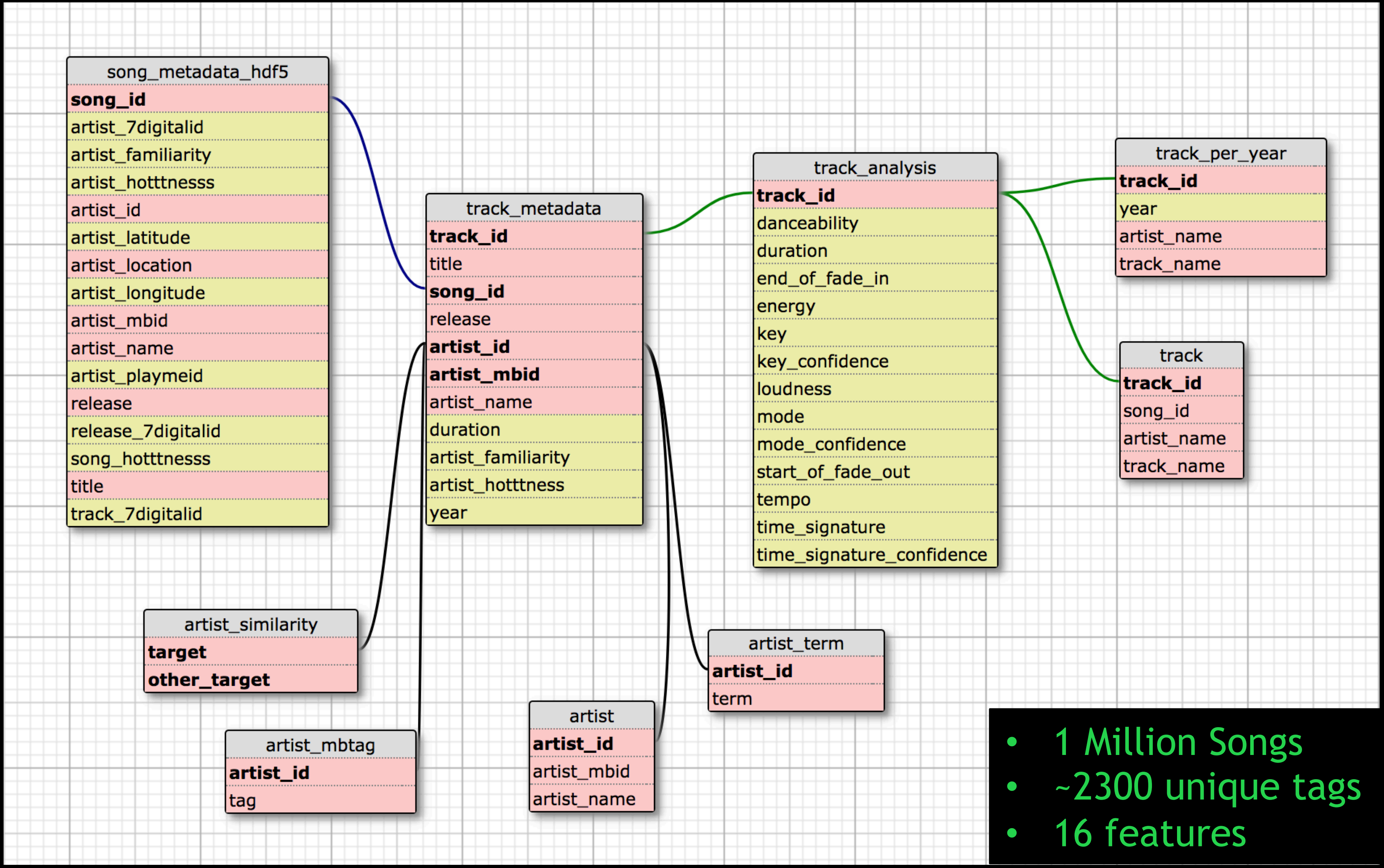
Genre Classification

Backstory

- Spotify noticed my investigation on their inner-workings and reached out asking for my humble advice on the next big innovation
- Consulting to determine the benefit of building out new social products on existing platform
- Million Song Dataset to help understand the use of social data to accompany current song data or lack of data

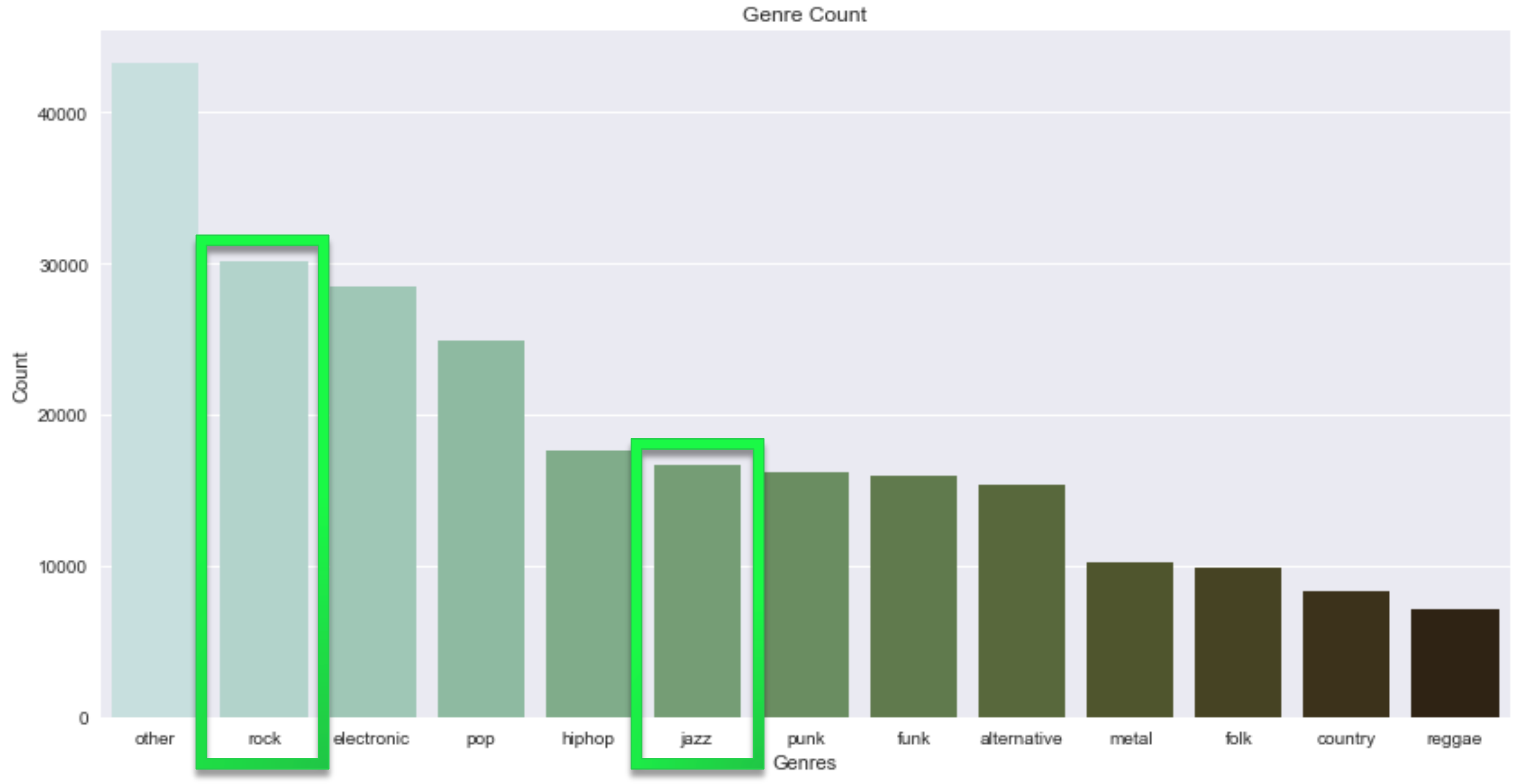
Million Songs Dataset



- 1 Million Songs
- ~2300 unique tags
- 16 features

Genres/ Labels

- 7,643 unique genres -> focused on 12 major genres
- Used Rock and Jazz for preliminary analysis

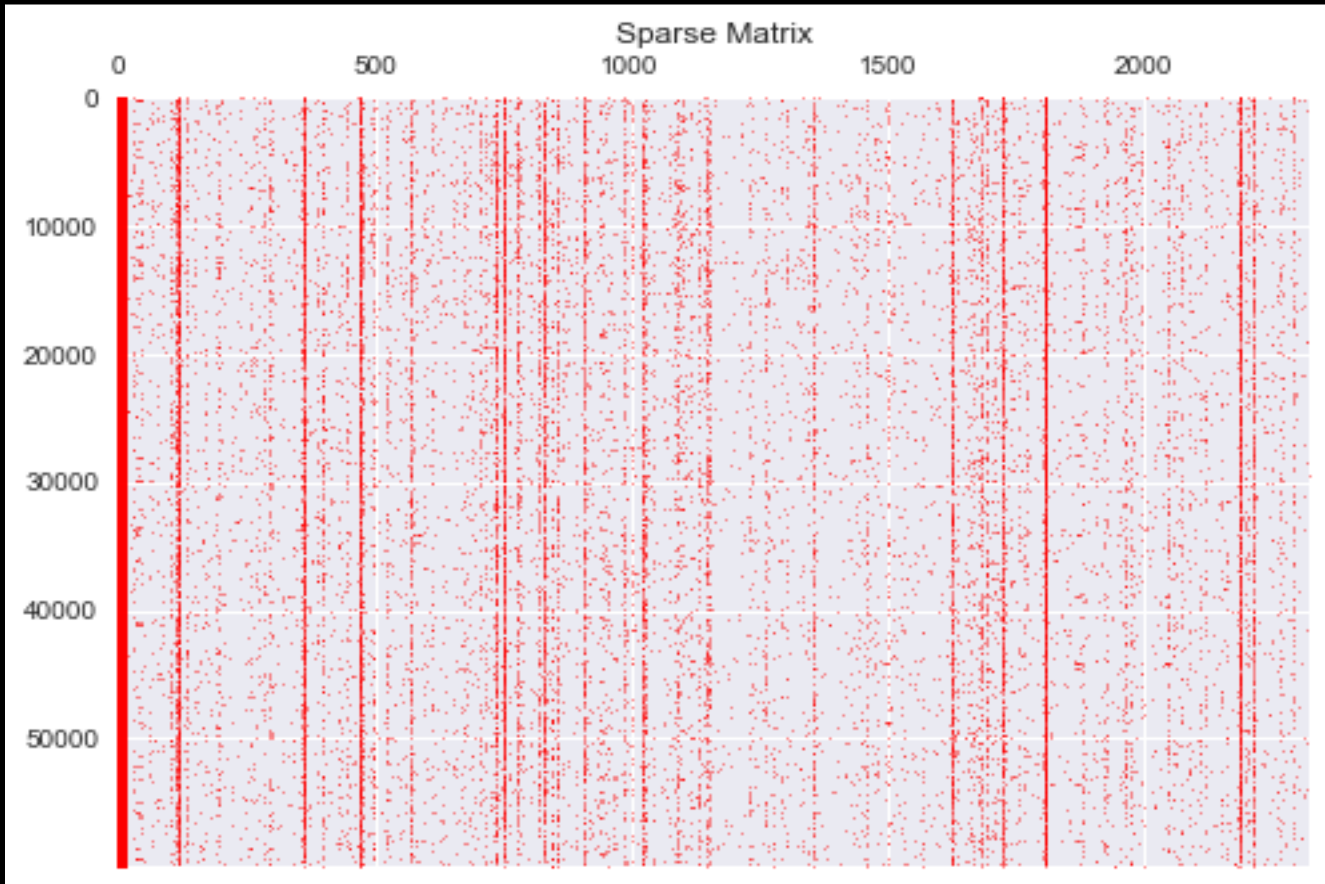


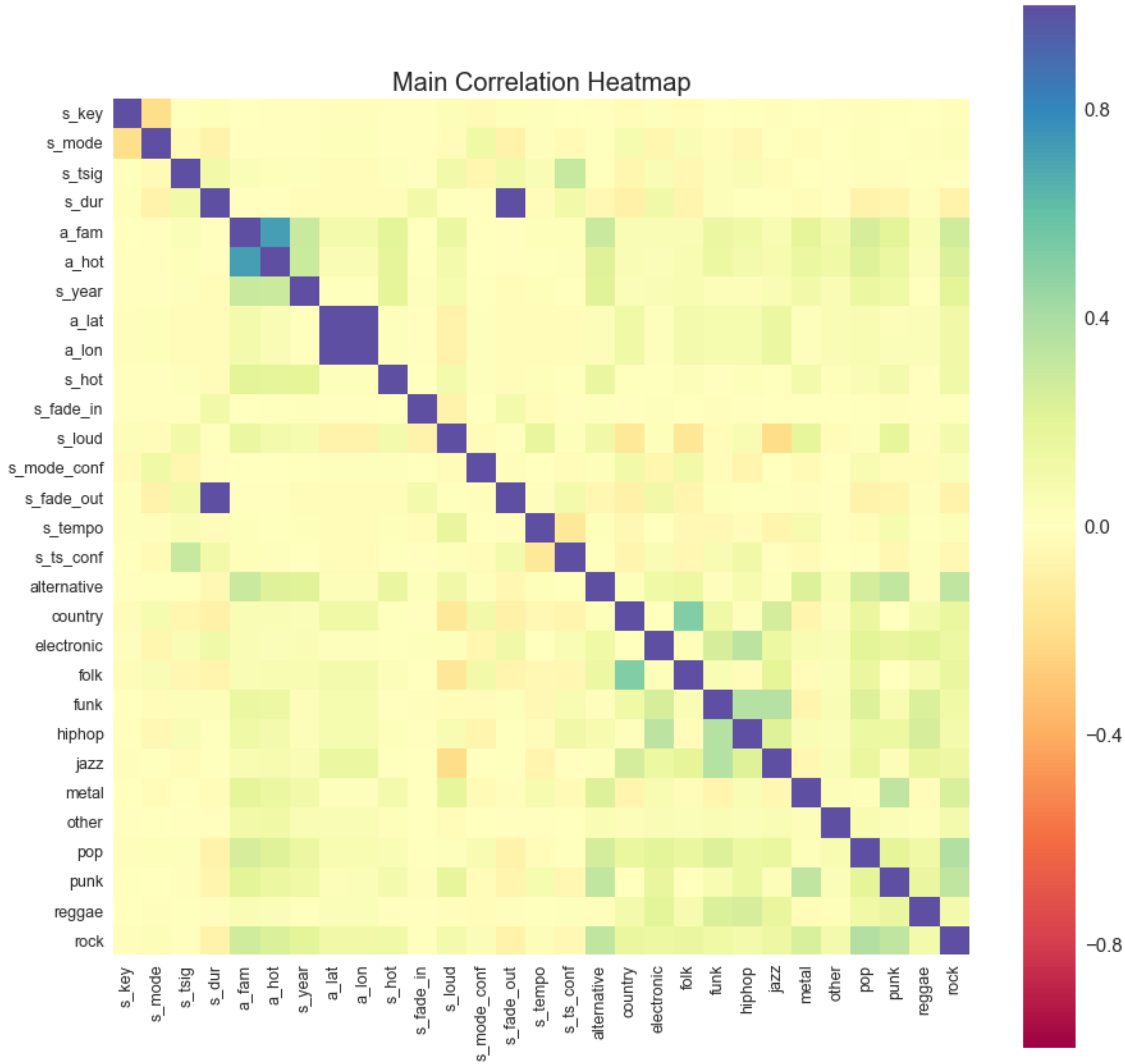
Tags

- ~2300 tags
- Nationality: British, American, French, German, World, Canadian, Latin, Italian, etc.

- Genres:
- Rock & Indie
- Hip-Hop/ RnB
- Punk

- Descriptive:
- Composer
- Female Vocal
- Soundtrack
- Musical





Main Correlation Heatmap
s_key
s_mode
s_tsig
s_dur
a_fam
a_hot
s_year
a_lat
a_lon
s_hot
s_fade_in
s_loud
s_mode_conf
s_fade_out
s_tempo
s_ts_conf
alternative
country
electronic
folk
funk
hiphop
jazz
metal
other
pop
punk
reggae
rock
0.8
0.4
0.0
−0.4
−0.8

Logistic Regression - No Tags

Odds Ratio

	2.5%	97.5%	OR
Intercept	0.744190	0.815180	0.778877
s_key	0.996968	0.999789	0.998377
s_mode	1.133105	1.157797	1.145385
s_tsig	0.990786	0.999410	0.995089
s_dur	1.007715	1.009202	1.008458
a_fam	25.235073	27.964123	26.564576
a_hot	2.744968	3.050567	2.893736
s_year	1.000325	1.000336	1.000331
a_lat	0.997497	0.997852	0.997675
a_lon	1.002228	1.002588	1.002408
s_hot	1.018527	1.020780	1.019653
s_fade_in	1.034574	1.040319	1.037443
s_loud	1.027170	1.029176	1.028173
s_mode_conf	1.568058	1.651688	1.609330
s_fade_out	0.989342	0.990825	0.990083
s_tempo	1.000809	1.001102	1.000956
s_ts_conf	0.731680	0.752952	0.742240

Logit Regression Resu

Dep. Variable:	rock	No. Obse	
Model:	Logit	Df Resid	
Method:	MLE	Df Model	
Date:	Tue, 20 Feb 2018	Pseudo R	
Time:	14:52:37	Log-Like	
converged:	True	LL-Null:	
		LLR p-va	

	coef	std err	z
Intercept	-0.2499	0.023	-10.751
s_key	-0.0016	0.001	-2.253
s_mode	0.1357	0.005	24.683
s_tsig	-0.0049	0.002	-2.227
s_dur	0.0084	0.000	22.395
a_fam	3.2796	0.026	125.192
a_hot	1.0625	0.027	39.458
s_year	0.0003	2.72e-06	121.343
a_lat	-0.0023	9.08e-05	-25.646
a_lon	0.0024	9.17e-05	26.223
s_hot	0.0195	0.001	34.523
s_fade_in	0.0368	0.001	26.022
s_loud	0.0278	0.000	55.823
s_mode_conf	0.4758	0.013	35.896
s_fade_out	-0.0100	0.000	-26.092
s_tempo	0.0010	7.48e-05	12.763
s_ts_conf	-0.2981	0.007	-40.773

Log Odds

- For Rock, Artist Familiarity score is very useful for classification
- Feature importance not consistent across genres

Scoring on Holdout-Test Set

No Tags Model	F1-score	ROC-AUC	Accuracy
Decision Tree			
Rock	.877	.669	.768
Jazz	.628	.719	.660
Logistic Regression			
Rock	.869	.685	.770
Jazz	.548	.650	.610
Logistic Regression - Balanced			
Rock	.764	.729	.676
Jazz	.582	.652	.609

Scoring on Holdout-Test Set

No Tags Model		F1-score	ROC-AUC	Accuracy
Decision Tree				
	Rock	.877	.669	.769
	Jazz	.628	.719	.660
Logistic Regression				
	Rock	.869	.685	.770
	Jazz	.548	.650	.610
Logistic Regression - Balanced				
	Rock	.764	.729	.676
	Jazz	.582	.652	.609

Scoring on Holdout-Test Set

Full Model		F1-score	ROC-AUC	Accuracy
Decision Tree				
	Rock	.869	.524	.769
	Jazz	.030	.511	.534
Logistic Regression				
	Rock	.871	.730	.776
	Jazz	.568	.676	.631
Logistic Regression - Balanced				
	Rock	.762	.775	.681
	Jazz	.602	.687	.635

Scoring on Holdout-Test Set

Full Model		F1-score	ROC-AUC	Accuracy
Decision Tree				
	Rock	.869	.524	.769
	Jazz	.030	.511	.534
Logistic Regression				
	Rock	$\boxed{.871}$	.730	$\boxed{.776}$
	Jazz	.568	.676	.631
Logistic Regression - Balanced				
	Rock	.762	$\boxed{.775}$	.681
	Jazz	.602	.687	.635

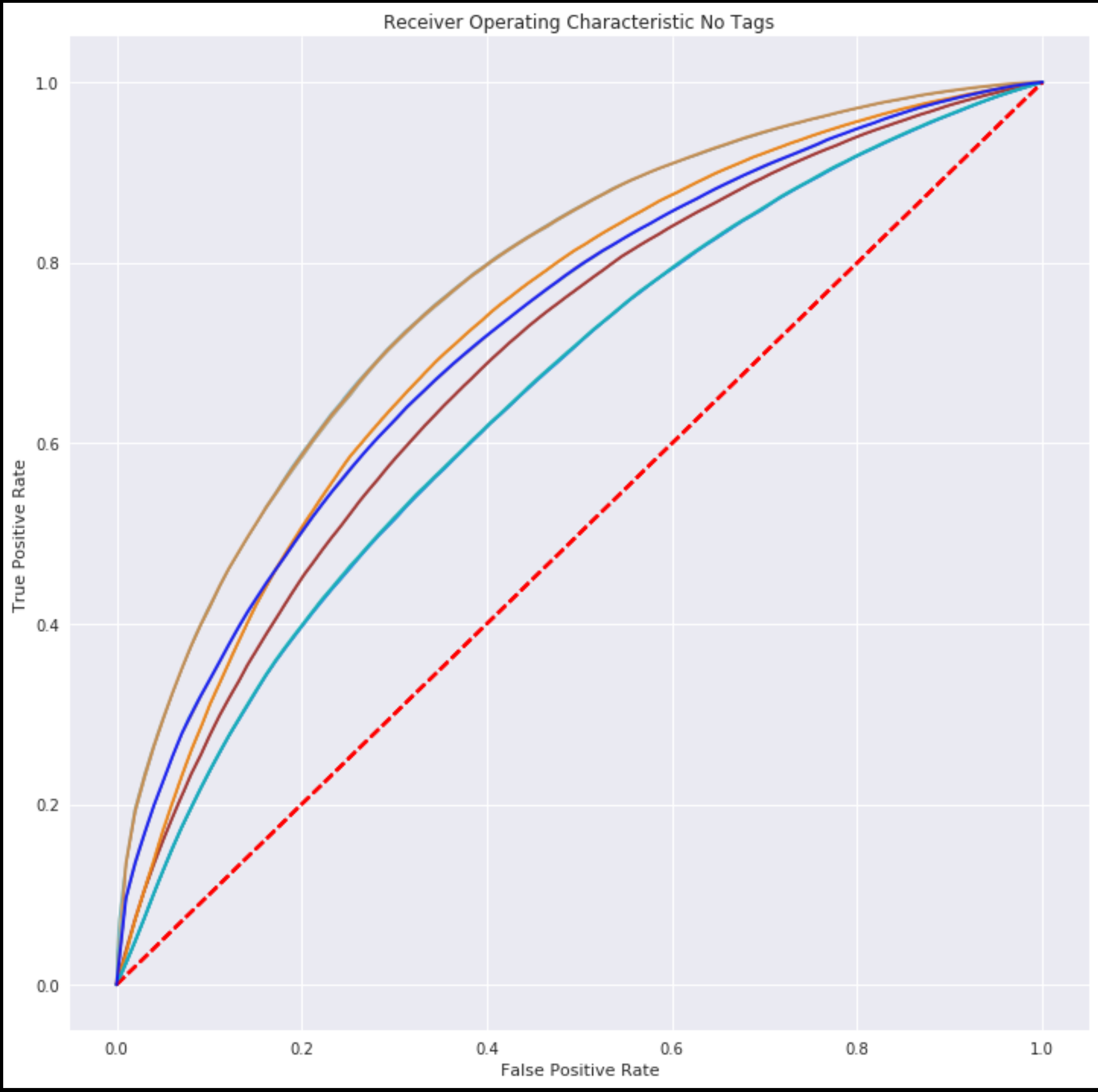
ROC - No Tags

Red/Orange – Rock

Blues – Jazz

Best Performer:
Rock –
Decision Tree
(Peru color)

Jazz –
Decision Tree (Dark Blue)



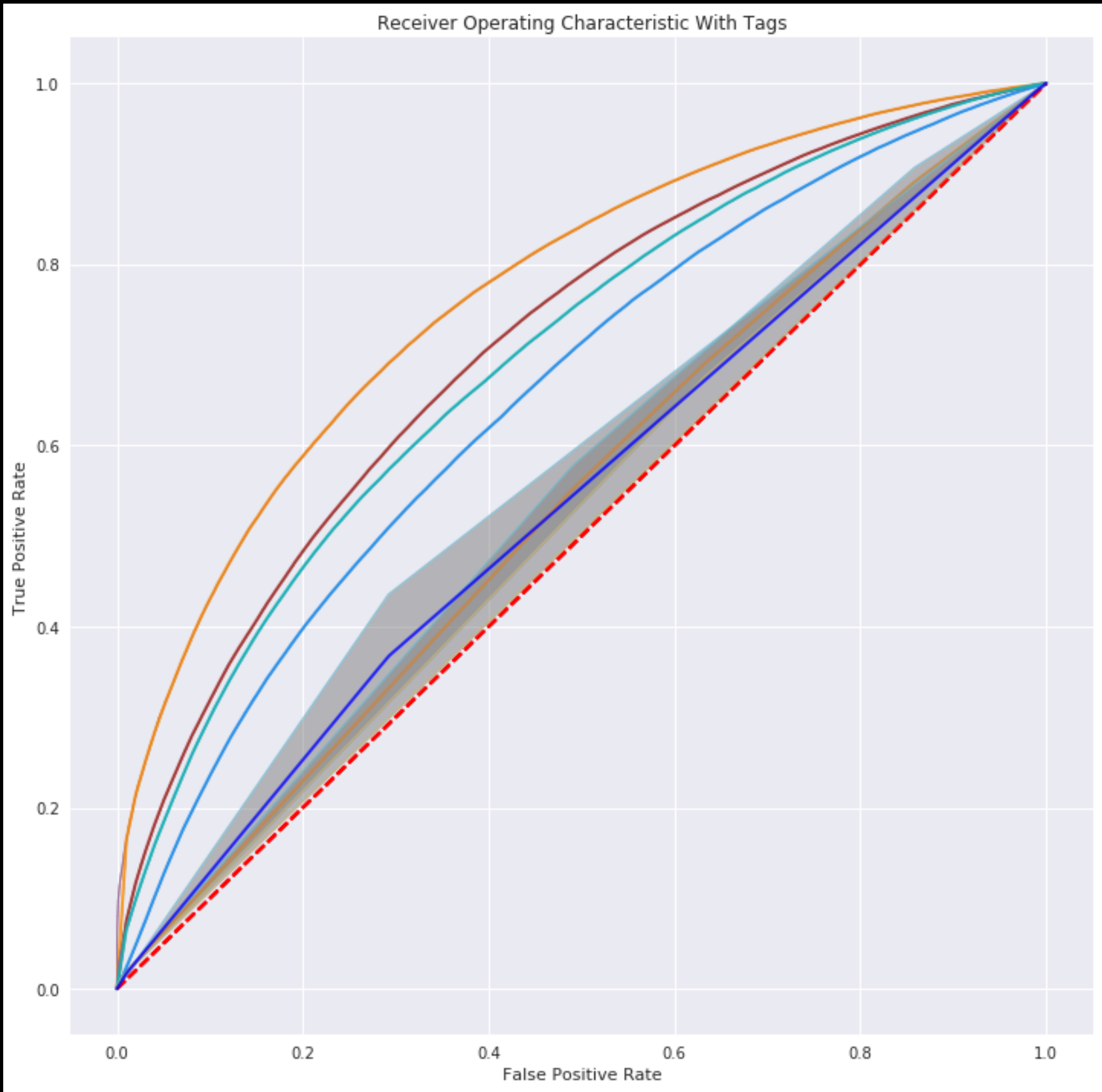
ROC – Tags

Red/Orange – Rock

Blues – Jazz

Best Performer:
Rock –
Balanced Logistic Regression (Orange)

Jazz –
Balanced Logistic Regression (Cyan)



Next Steps

- Random Forest would be nice to use to address curse of dimensionality
- Further data collection in other areas would also be beneficial (better song description, lyrics, song comments, etc.)
- Look more deeply at features, their individual contribution and see what category of tag works best for improving models
- More visualizations of data
- Compare a tag only dataset to main features

Conclusion

- The gains from adding in user tags are noticeable up to 6% improvement in AUC scoring and with a little more effort could significantly help our models for classifying music
- Through the wisdom of crowds, we're able to overcome the cold start problem (lack of data)
- We can find good signal in user generated tags, using filters on bad actors, should we lack user behavior
- Social data not an urgent set of features, but should look into a long term plan of adding in more social aspects into the platform
- Data cleaning and maintenance efforts would probably deem more useful and should be addressed first!