Spotify - Predicting Song Popularity

Bader Aldawoodi, Gilbert Arellano, Cole Matsueda, Austin Simpson

motivation/business value

Improved songwriting: By understanding which factors are most important for predicting the popularity of a song, artists can use this information to improve their songwriting and create songs that are more likely to become popular.

Targeted marketing: By knowing which songs are likely to become popular, artists can target their marketing efforts more effectively, promoting their most popular songs to the right audiences at the right time.

Increased revenue: By promoting their most popular songs, artists can increase the number of plays for those songs, leading to increased revenue through advertising and subscriptions.

Overall, the information provided by a model that can predict the popularity of a Spotify song can be incredibly valuable for artists on the platform, helping them improve their songwriting, target their marketing efforts, and increase their revenue.

Dataset Variables

- Artist
- Album
- Track_number
- Id
- Name_x
- Uri_x
- Acousticness
- Danceability
- Energy
- Instrumentalness
- Liveness
- Loudness
- Speechiness
- Tempo
- valence
- Popularity
- Duration_ms
- Explicit
- Mode
- Time_signature
- Genres
- Artist_pop
- uri_y

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
track_number	7534	8.46	5.867	1	4	12	50
acousticness	7534	0.213	0.259	0	0.016	0.322	0.995
danceability	7534	0.602	0.178	0	0.481	0.742	0.966
energy	7534	0.655	0.193	0.003	0.529	0.802	0.998
instrumentalness	7534	0.109	0.253	0	0	0.017	0.997
liveness	7534	0.229	0.207	0	0.099	0.3	0.995
loudness	7534	-7.382	3.276	-31.909	-8.621	-5.346	-0.674
speechiness	7534	0.106	0.12	0	0.04	0.117	0.96
tempo	7534	119.544	28.068	0	98.032	136.017	218.365
valence	7534	0.473	0.246	0	0.277	0.661	0.979
popularity	7534	30.927	20.962	0	14	46	96
duration_ms	7534	227855.453	89558.782	8000	185688	256473.25	3174106
explicit	7534						
... No	6363	84.5%					
... Yes	1171	15.5%					
mode	7534	0.608	0.488	0	0	1	1
time_signature	7534	3.921	0.38	0	4	4	5
artist_pop	7534	76.882	12.941	28	68	85	97
total_followers	7534	16762675.861	18891900.787	15585	2658308	24922915	68861670

Data cleaning

Removing duplicate songs

Some columns were duplicates such as name_y and artist

Removed genres because it was in an unusable format

```
data <- read_csv("all_data.csv")

data_clean = data %>% select(-name_x, -uri_x, -uri_y, -album, -id,
-name_y, -genres)

data_clean = data_clean %>% as_tibble() %>%
  mutate(explicit = as.factor(explicit),
         mode = as.factor(mode),
         time_signature = as.factor(time_signature),
         artist = as.factor(artist)) %>%
  mutate_if(is.character, as.factor)
```

Linear Regression Variables

Continuous:

- Total_followers
- Energy
- Liveness
- Loudness
- Instrumentalness
- Tempo
- Valence
- Danceability
- Speechiness
- Acousticness
- duration_ms

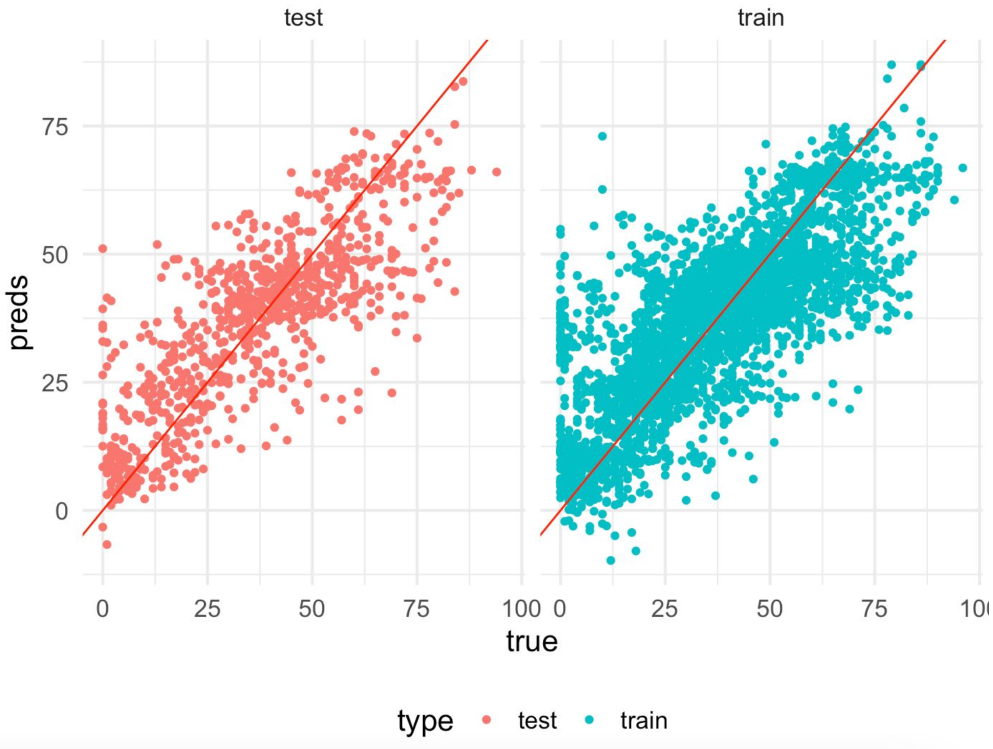
Categorical:

- Artist
- Explicit
- time_signature

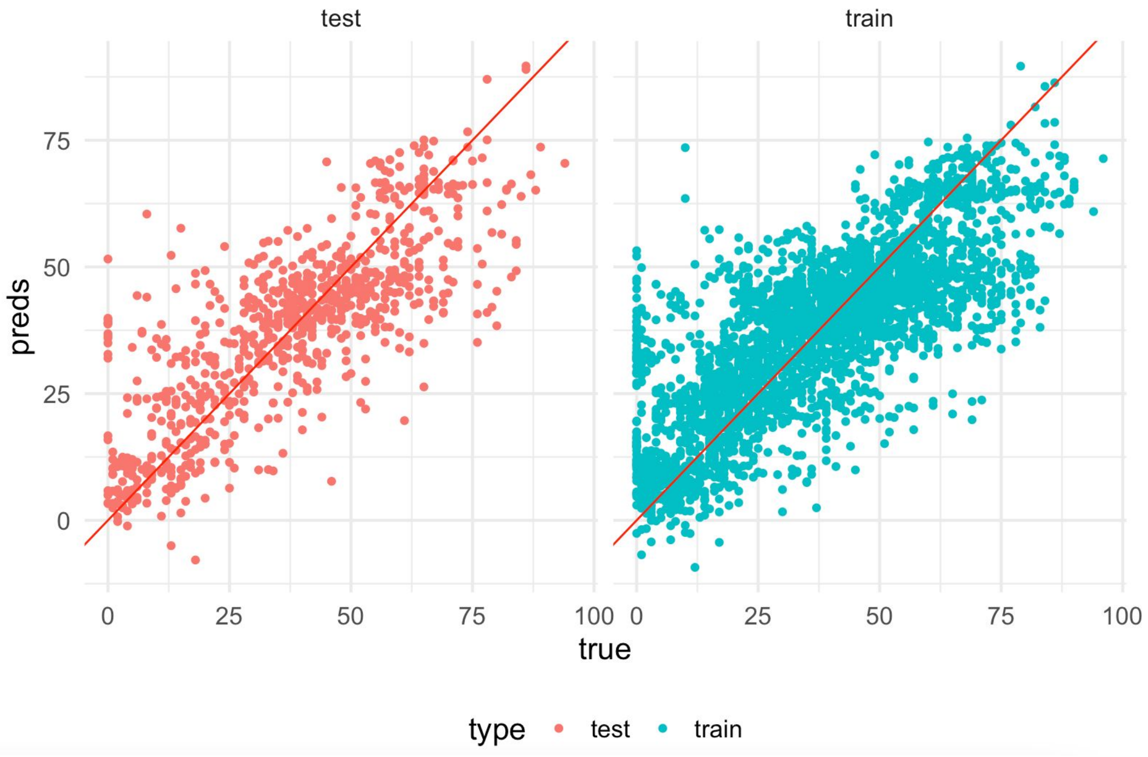
Linear Regression

```
spotify <- dummy_cols(spotify, select_columns = "artist")
spotify <- dummy_cols(spotify, select_columns = "explicit")
spotify <- dummy_cols(spotify, select_columns = "time_signature")

spotify_split <- initial_split(spotify, prop = 0.8)
spotify_train <- training(spotify_split)
spotify_test <- testing(spotify_split)

#with artists
mod <- lm(popularity ~ total_followers + energy + liveness + loudness
          + instrumentalness + tempo + valence + danceability + speechiness
          + acousticness + duration_ms + artist + explicit + time_signature, data = spotify_train)

#without artists
mod2 <- lm(popularity ~ total_followers + energy + liveness + loudness
          + instrumentalness + tempo + valence + danceability + speechiness
          + acousticness + duration_ms + explicit + time_signature, data = spotify_train)
```

Linear Regression with Artists



Linear Regression without Artists



Linear Regression

```r
getrmse <- function(true, predictions){
  sqrt(mean((true - predictions)^2))
}

getrmse(spotify_train$popularity, preds_train2)

getrmse(spotify_test$popularity, preds_test2)
```

```
> getrmse <- function(true, predictions){
+   sqrt(mean((true - predictions)^2))
+ }
>
> getrmse(spotify_train$popularity, preds_train2)
[1] 16.73913
>
> getrmse(spotify_test$popularity, preds_test2)
[1] 16.81176
```

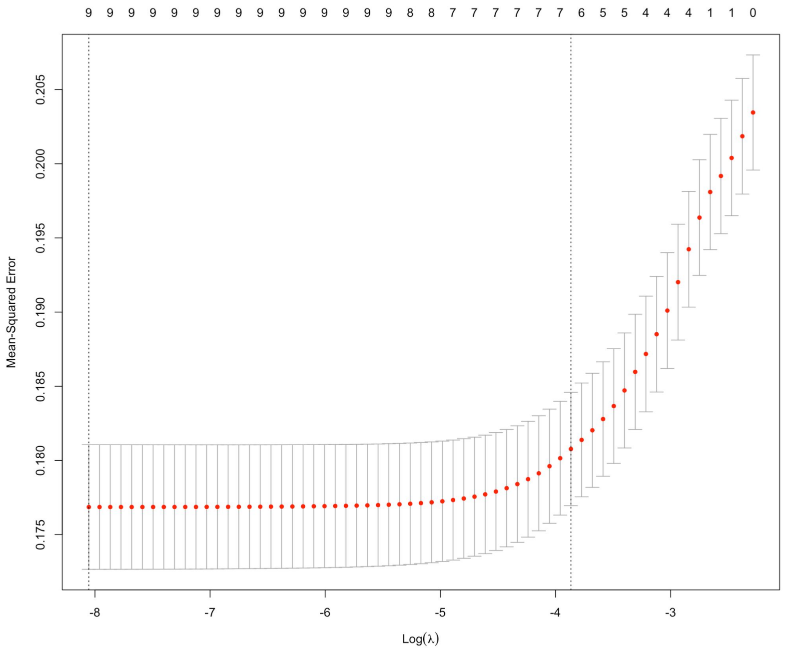
Linear Regression Summary with Artists

	popularity		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	15313.75	-2940278.88 – 2970906.38	0.992
total followers	-0.00	-0.25 – 0.24	0.992
energy	-4.09	-7.49 – -0.69	**0.019**
liveness	-7.90	-9.88 – -5.92	**<0.001**
loudness	0.52	0.33 – 0.71	**<0.001**
instrumentalness	-6.49	-8.37 – -4.61	**<0.001**
tempo	0.01	-0.00 – 0.03	0.123
valence	-0.61	-2.90 – 1.68	0.602
danceability	7.43	4.02 – 10.84	**<0.001**
speechiness	-8.82	-12.49 – -5.15	**<0.001**
acousticness	0.17	-1.84 – 2.18	0.868
duration ms	-0.00	-0.00 – 0.00	0.418
artist [AJR]	-11892.32	-2314804.50 – 2291019.86	0.992
artist [Anderson .Paak]	-12419.59	-2414650.71 – 2389811.54	0.992
artist [Anomalie]	-15141.06	-2943077.32 – 2912795.20	0.992
artist [Beyoncé]	21554.00	-4146037.59 – 4189145.59	0.992
artist [Birocratic]	-15233.23	-2959120.35 – 2928653.90	0.992
artist [Black Pistol Fire]	-15041.96	-2923222.56 – 2893138.65	0.992
artist [Brasstracks]	-15157.98	-2944800.39 – 2914484.42	0.992
artist [Britney Spears]	-2884.09	-561444.12 – 555675.95	0.992
artist [Calvin Harris]	13892.91	-2674034.52 – 2701820.34	0.992
artist [Conro]	-15225.63	-2958459.81 – 2928008.54	0.992
artist [Darondo]	-15158.35	-2946480.63 – 2916163.94	0.992
artist [Doja Cat]	-1924.70	-377709.22 – 373859.82	0.992
artist [Drake]	72136.58	-13879516.99 – 14023790.14	0.992
artist [EMEFE]	-15281.14	-2967047.28 – 2936485.01	0.992
artist [FKJ]	-13686.77	-2664315.56 – 2636942.03	0.992
artist [Foo Fighters]	-2085.94	-406360.17 – 402188.29	0.992
artist [Imagine Dragons]	40143.26	-7722515.92 – 7802802.43	0.992
artist [James Brown]	-13016.50	-2525610.77 – 2499577.76	0.992
artist [Just A Gent]	-15132.45	-2941505.69 – 2911240.80	0.992
artist [Lil Nas X]	76.23	-11389.03 – 11541.49	0.990
artist [Louis The Child]	-14726.76	-2859833.77 – 2830380.24	0.992
artist [Magic City Hippies]	-15121.97	-2939885.80 – 2909641.86	0.992
artist [Maroon 5]	33860.99	-6516456.98 – 6584178.97	0.992
artist [Michael Jackson]	16359.23	-3147241.28 – 3179959.74	0.992
artist [Muse]	-6106.64	-1189413.07 – 1177199.78	0.992
artist [My Chemical Romance]	-7337.75	-1428609.26 – 1413933.75	0.992
artist [Olivia Rodrigo]	12946.86	-2484172.70 – 2510066.42	0.992
artist [OneRepublic]	3261.91	-626283.74 – 632807.57	0.992
artist [Royal Blood]	-13388.81	-2606071.05 – 2579293.44	0.992
artist [Sabrina Carpenter]	-9341.65	-1818709.01 – 1800025.71	0.992
artist [Snarky Puppy]	-14512.53	-2821511.62 – 2792486.55	0.992
artist [Stevie Wonder]	-7829.82	-1519867.89 – 1504208.26	0.992
artist [Still Woozy]	-14155.60	-2755169.18 – 2726857.98	0.992
artist [Tee Grizzley]	-11342.37	-2203802.28 – 2181117.54	0.992
artist [The Black Keys]	-10489.84	-2040020.73 – 2019041.06	0.992
artist [The Weeknd]	51553.47	-9916800.38 – 10019907.31	0.992
artist [Travis Scott]	11845.91	-2276865.35 – 2300557.17	0.992
artist [Vulfpeck]	-14299.25	-2781520.39 – 2752921.88	0.992
artist [Young the Giant]	-13585.29	-2641418.09 – 2614247.51	0.992
artist [Zedd]	-7862.41	-1527525.02 – 1511800.19	0.992
explicitTRUE	9.44	7.69 – 11.19	**<0.001**
time signature	1.48	0.42 – 2.53	**0.006**
Observations	4154		
R^2 / R^2 adjusted	0.614 / 0.609		

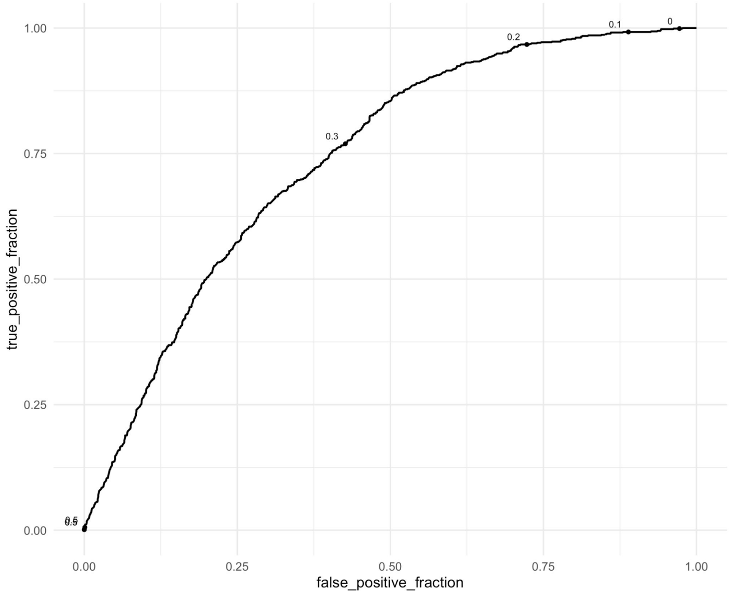
Linear Regression without Artists

	popularity		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	35.44	27.59 – 43.30	**<0.001**
total followers	0.00	0.00 – 0.00	**<0.001**
energy	-5.26	-9.81 – -0.71	**0.023**
liveness	-7.71	-10.44 – -4.99	**<0.001**
loudness	0.74	0.50 – 0.99	**<0.001**
instrumentalness	-7.85	-10.13 – -5.57	**<0.001**
tempo	0.02	-0.00 – 0.04	0.056
valence	-9.28	-12.14 – -6.43	**<0.001**
danceability	-6.48	-10.63 – -2.34	**0.002**
speechiness	-13.23	-18.08 – -8.38	**<0.001**
acousticness	-5.33	-7.98 – -2.67	**<0.001**
duration ms	-0.00	-0.00 – 0.00	0.059
explicitTRUE	13.40	11.73 – 15.08	**<0.001**
time signature	3.77	2.30 – 5.24	**<0.001**
Observations	3323		
R^2 / R^2 adjusted	0.377 / 0.375		

Lasso Model



9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 8 8 7 7 7 7 7 7 6 5 5 4 4 4 1 1 0
Mean-Squared Error
0.175
0.180
0.185
0.190
0.195
0.200
0.205
-8
-7
-6
-5
-4
-3
Log(λ)



true_positive_fraction
false_positive_fraction
0.00
0.25
0.50
0.75
1.00
0.3
0.2
0.1
0

Lasso Model

```r
for (x in incrementalTreshold) {

  # Mutate isPopular variable with a new threshold, x
  spotify_clean_1 <- spotify %>%
  mutate(
    isPopular = if_else(popularity >= x, 1, 0),
  )
```

(Intercept)	0.638
danceability	0.267
energy	-0.266
instrumentalness	-0.301
liveness	-0.211
loudness	0.024
speechiness	0.036
tempo	0.001
valence	-0.362
duration_ms	0.000

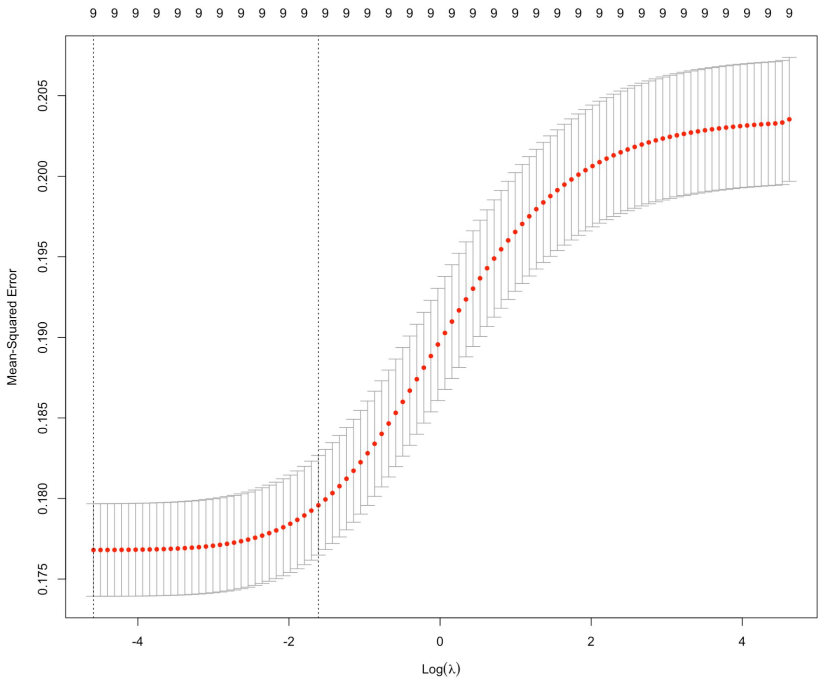
Error Metric Measurements for Lasso

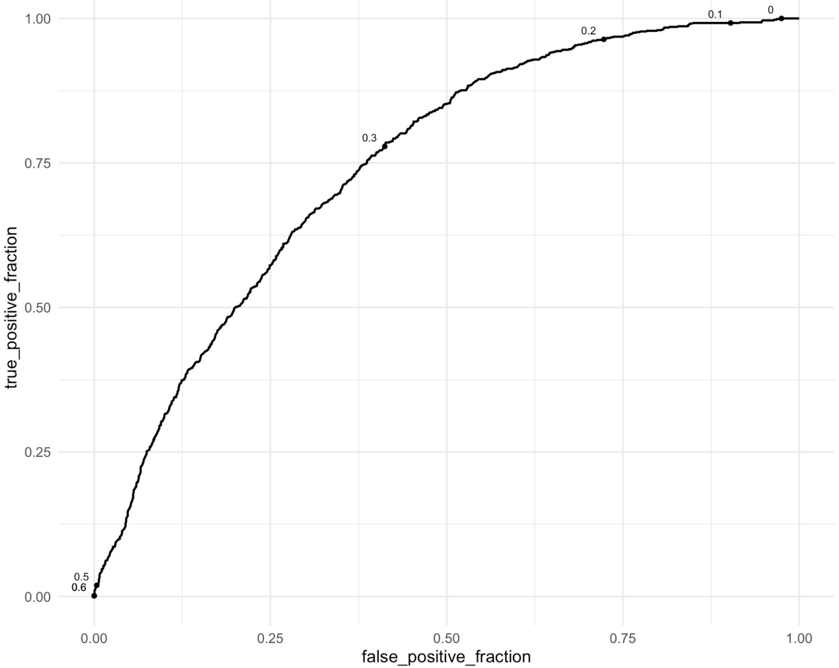
Deciding which isPopular threshold is most accurate

ISPOPULAR	ROC_AUC	Λ
50	0.500	-8.057
55	0.500	-6.045
65	0.500	-6.767
70	0.500	-9.278
75	0.500	-9.709

```r
lasso_fit_1 <- cv.glmnet(isPopular ~ danceability + energy + instrumentalness
                   + liveness + loudness + speechiness + tempo + valence + duration_ms,
                   data = spotify_train_1,
```

Ridge Model


9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
Mean-Squared Error
0.175
0.180
0.185
0.190
0.195
0.200
0.205
-4
-2
0
2
4
Log(λ)


true_positive_fraction
false_positive_fraction
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
0
0.1
0.2
0.3
0.5
0.6

Ridge Model

(Intercept)	0.485
danceability	0.164
energy	-0.122
instrumentalness	-0.228
liveness	-0.174
loudness	0.014
speechiness	0.017
tempo	0.001
valence	-0.231
duration_ms	0.000

Error Metric Measurements for Ridge

Deciding which isPopular threshold is most accurate

ISPOPULAR	ROC_AUC	Λ
50	0.743	-4.571
55	0.750	-4.760
65	0.761	-5.384
70	0.500	-4.216
75	0.500	-4.490

```
ridge_fit_1 <- cv.glmnet(isPopular ~ danceability + energy + instrumentalness
                  + liveness + loudness + speechiness + tempo + valence + duration_ms,
                  data = spotify_train_1,
```

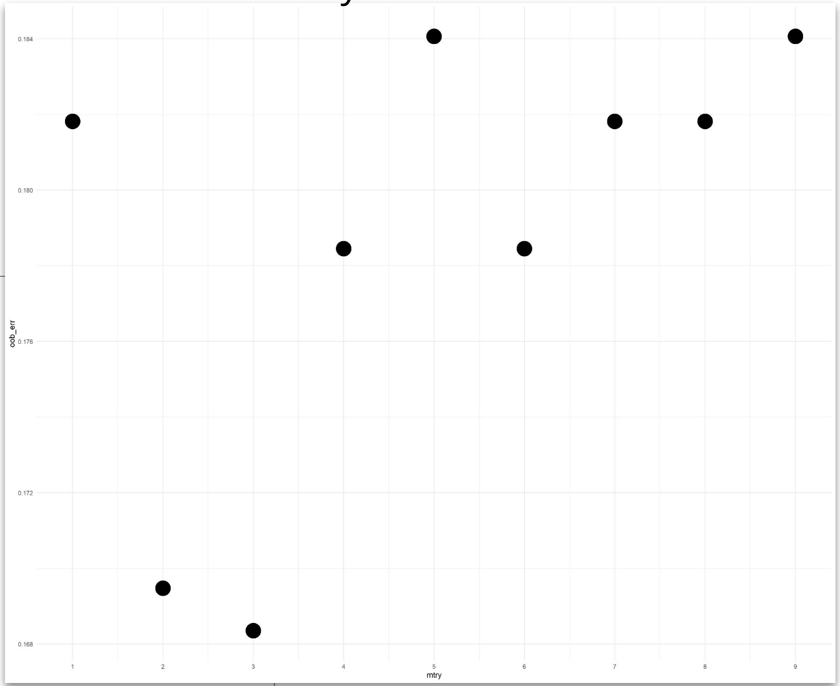
Random Forest

Hyperparameter Tuning

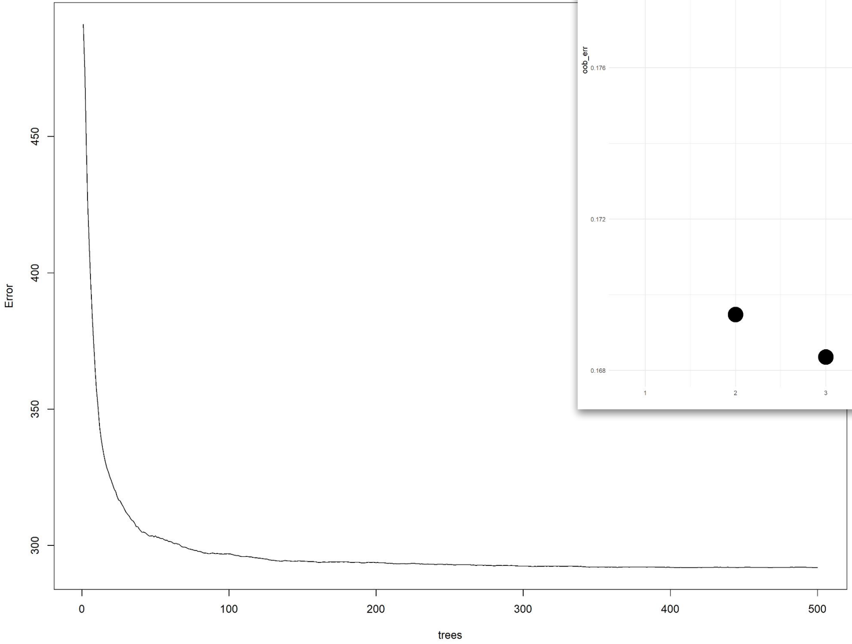
ntree = 200

mtry = 3

Mtry vs error

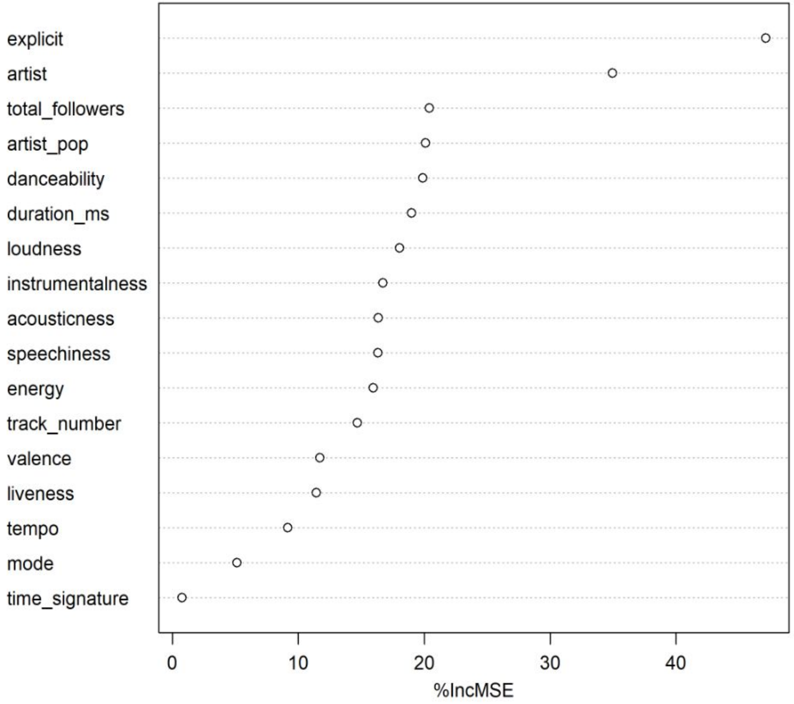


Ntree vs error

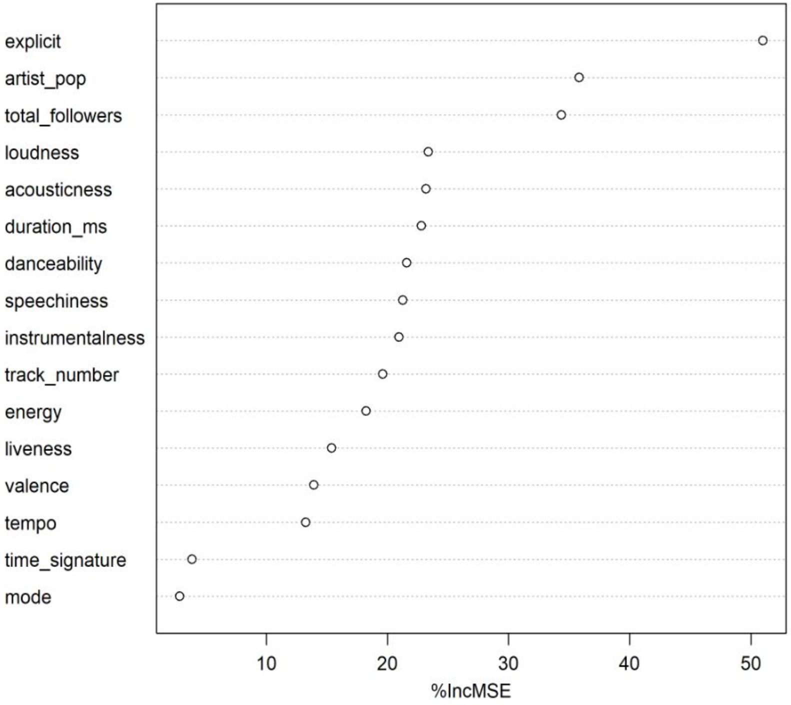


Variable Importance

With artist



Without artist



Random Forest

R Squared with artist

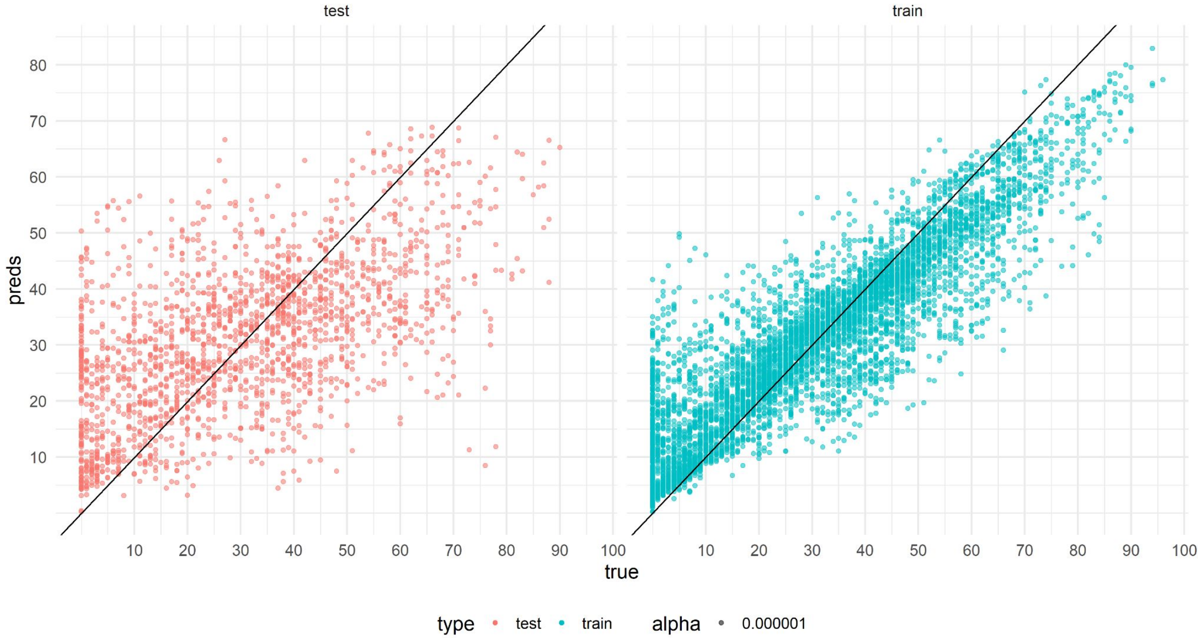
Train: 0.8034109

Test: 0.3316403

R Squared without artist

Train: 0.8040931

Test: 0.3038147



Conclusion

- **Linear Regression:** Using this data set, we can conclude that our model is not a good model in predicting the popularity of a song since the R2 score is very low and the model doesn't fit on the dataset very well.
- **Regularization:** This model had a poor ROC/AUC and R^2 Score and was unable to predict the popularity
- **Random Forests:** This model was very overfit and the R2 Score on the test set was only 0.33

Github Link: https://github.com/arell110/MGSC_310_FALL_2022: