Spotify Song Recommendation Model

Alex Shen, Matt Robbins

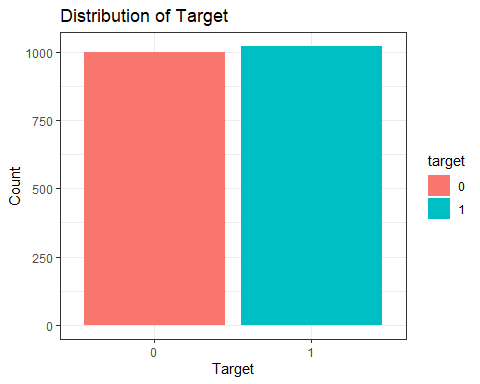
## Introduction

Spotify is a music streaming service with 356 million monthly active users and over 70 million tracks in their library. What drives their success? Aside from the size of the library, it is their ability to keep users engaged by providing a constant stream of relevant song and even podcast recommendations. In order to make these recommendations, there must be some data on songs as well as user preference data that can help decide which songs to recommend to a particular person. Some of Spotify’s competitors such as Pandora will tag songs with attributes manually, while Spotify has the advantage of using deep learning models to do this as well as synthesize this with artist, genre, and user preference information in order to better inform recommendations. The three main recommendation models used at Spotify include collaborative filtering (which analyzes similarity in behavior between users), natural language processing or NLP (which extracts information from words in song lyrics, song titles, and playlist titles), and audio models (which detect patterns in raw audio content). (Boyd)

Our data set came from Kaggle, and was based on one person’s music preferences. Each observation in the data is a song on Spotify, and we are given information about each song, as well as whether or not the person liked the song. Continuous variables of interest that are between 0 and 1 include acousticness, danceability, energy, instrumentalness (amount of vocals in a song), liveness (how likely the song was recorded live), loudness, speechiness (amount of spoken words in a song), and valence (measure of how happy a song sounds). Other variables of interest include duration\_ms (continuous variable denoting duration of the song in ms), key (categorical variable denoting the key of the song from 0-11), mode (modality of the song expressed as 0 or 1, corresponding to major or minor), tempo (continuous variable in BPM), time\_signature (categorical variable from 1-5), and target (whether or not the person liked the song expressed as a 0 or 1). Most of these variables come from Spotify’s automated deep learning models that processes the raw audio of songs.

We wish to do a Bayesian logistic regression using the information about the song as predictors and target as the response variable. With such a model, the best use of it would be as a song recommender algorithm. We would be able to predict whether or not the person likes a song, and if we predict that they like the song, we would recommend it. One limitation of using this method to recommend a song is that it requires user input. Most users will not take the time to record whether or not they like or dislike songs that they come across, which is why Spotify’s recommendation model takes into account factors such as time a song is played for or if a person has the song in a playlist in order to determine if a person likes a song or not. Despite our proposed method not being useful as a recommendation algorithm for Spotify itself, we do have the upside of being able to definitively know whether or not a person likes a song, rather than having to infer on it as Spotify does. So, for people who are fine with compiling a list of songs that they explicitly like or dislike, our model should be able to make a more targeted recommendation.

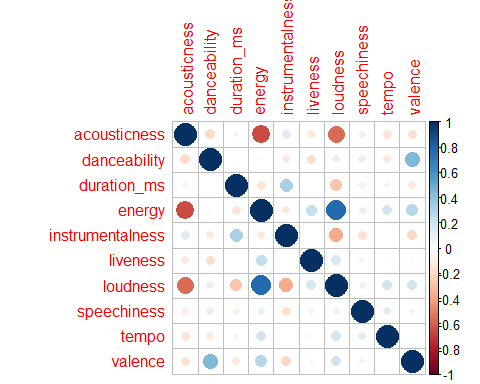
## Data



The data set has 2017 songs. We decided to randomly divide this data, with 75% of it becoming the data on which we will train the model on, and the remaining 25% becoming data we will test the model on after it is created. We also check the distribution of target, since we want to make sure there are enough liked and disliked songs in the data set to be able to properly train our model. For example, we do not want to see only a handful of liked songs and almost all disliked songs, as this could result in a model that predicts almost every song as disliked, and will end up with a inflated correct prediction rate. However, as we can see, the data set is split almost into half liked songs and half disliked songs, which is beneficial for training our model.

After selecting the variables relevant to our analysis, we also factorized the categorical variables so that R would not accidentally treat them as continuous.

For our purposes, since we are trying to model and predict whether or not this person likes a given song, the response variable target takes the value of 0 or 1. That means our sampling model is a Bernoulli distribution with probability , or .



Because our continuous predictors are on different scales, we decided to normalize them. We then plot a correlation plot between each of the continuous predictors. As we can see, several variables have a large of correlation. Energy and acousticness as well as loudness and acousticness are negatively correlated, while energy and loudness are positively correlated. We are trying to avoid multicollinearity in our model, and since high correlation often is associated with multicollinearity, we will keep an eye on the pairs of variables with high magnitudes of correlation.

## Model Creation

We were not able to find another logistic regression on a data set of songs on Spotify in order to predict whether or not someone likes a song. As a result, we did not have much information about what the coefficients might be, and decided to choose weakly informative priors. In this project, we consider three different prediction models: a logistic regression model which uses Cauchy(0,2.5) priors for its coefficients, a logistic regression model which uses regularized horseshoe priors for its coefficients, and a “null” model which produces 1 (a “yes” for liking the song), regardless of input variables. We choose a Cauchy(0,2.5) prior for our first model because it is an empirically successful choice for weakly informative priors (Gelman, et al.). Regularized horseshoe priors are typically meant for models in which the number of predictive variables is large in comparison to the number of samples. Although this is not the case in our experiment (, ), we were still interested to see whether this model could outperform our standard model.

By specifying our prior distributions, will fit a logistic model to the data by MCMC sampling over the posterior distributions. We used the default rstanarm settings of running 4 Markov chains for each distribution, with each chain being 2000 iterations. As shown below, using a Rhat convergence diagnostic for each variable in each model, we find that all variables produced a Rhat value less than 1.05, which is within the valid range for convergence of the MCMC process (Guo, et al.).

We decided to backwards select our models using Bayes factors, which are useful in selecting Bayesian models. In our full model, we included all predictors with target as the response variable. We also decided to add some interaction effects. We thought that the mode variable (modality of a song, major or minor) would affect the danceability of a song as well as the valence (measurement of how happy a song sounds). We also thought that the time signature of a song might affect its tempo. As a result, we added interactions between mode and danceability, mode and valence, and time signature and tempo.

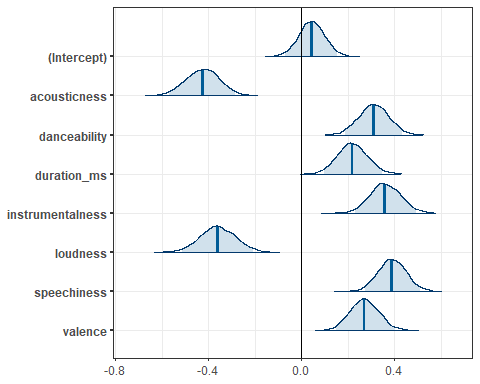
We calculated the Bayes factor of each coefficient compared to a point-null, which is the Savage-Dickey Ratio (Makowski, et al.). We also checked the p value of a hypothesis test for each coefficient equaling 0. We dropped the term with the highest p value and lowest Bayes factor (these were always in agreement). We stopped when each term had a Bayes factor that can be interpreted as strong (or better) support for the coefficient not equaling zero. (Lee & Wagenmakers)

### Model with Cauchy Prior

## Sampling priors, please wait...

## Bayes Factor (Savage-Dickey density ratio)  
##   
## Parameter | BF  
## -------------------------  
## (Intercept) | 0.025  
## acousticness | > 1000  
## danceability | 620.65  
## duration\_ms | 4.99  
## instrumentalness | 206.50  
## loudness | 137.48  
## speechiness | > 1000  
## valence | 23.30  
##   
## \* Evidence Against The Null: 0

## Call:  
## BF.glm(x = model)  
##   
## Bayesian hypothesis test  
## Type: exploratory  
## Object: stanreg  
## Parameter: general parameters  
## Method: adjusted fractional Bayes factors using Gaussian approximations  
##   
## Posterior probabilities:  
## Pr(=0) Pr(<0) Pr(>0)  
## (Intercept) 0.935 0.014 0.051  
## acousticness 0.000 1.000 0.000  
## danceability 0.000 0.000 1.000  
## duration\_ms 0.055 0.000 0.944  
## instrumentalness 0.000 0.000 1.000  
## loudness 0.000 1.000 0.000  
## speechiness 0.000 0.000 1.000  
## valence 0.002 0.000 0.998



##   
## Model Info:  
## function: stan\_glm  
## family: binomial [logit]  
## formula: target ~ acousticness + danceability + duration\_ms + instrumentalness +   
## loudness + speechiness + valence  
## algorithm: sampling  
## sample: 4000 (posterior sample size)  
## priors: see help('prior\_summary')  
## observations: 1513  
## predictors: 8  
##   
## Estimates:  
## mean sd 10% 50% 90%  
## (Intercept) 0.0 0.1 0.0 0.0 0.1   
## acousticness -0.4 0.1 -0.5 -0.4 -0.3   
## danceability 0.3 0.1 0.2 0.3 0.4   
## duration\_ms 0.2 0.1 0.1 0.2 0.3   
## instrumentalness 0.4 0.1 0.3 0.4 0.4   
## loudness -0.4 0.1 -0.5 -0.4 -0.3   
## speechiness 0.4 0.1 0.3 0.4 0.5   
## valence 0.3 0.1 0.2 0.3 0.4   
##   
## Fit Diagnostics:  
## mean sd 10% 50% 90%  
## mean\_PPD 0.5 0.0 0.5 0.5 0.5   
##   
## The mean\_ppd is the sample average posterior predictive distribution of the outcome variable (for details see help('summary.stanreg')).  
##   
## MCMC diagnostics  
## mcse Rhat n\_eff  
## (Intercept) 0.0 1.0 5101   
## acousticness 0.0 1.0 5508   
## danceability 0.0 1.0 5230   
## duration\_ms 0.0 1.0 5787   
## instrumentalness 0.0 1.0 6821   
## loudness 0.0 1.0 6378   
## speechiness 0.0 1.0 5373   
## valence 0.0 1.0 4442   
## mean\_PPD 0.0 1.0 4500   
## log-posterior 0.0 1.0 1904   
##   
## For each parameter, mcse is Monte Carlo standard error, n\_eff is a crude measure of effective sample size, and Rhat is the potential scale reduction factor on split chains (at convergence Rhat=1).

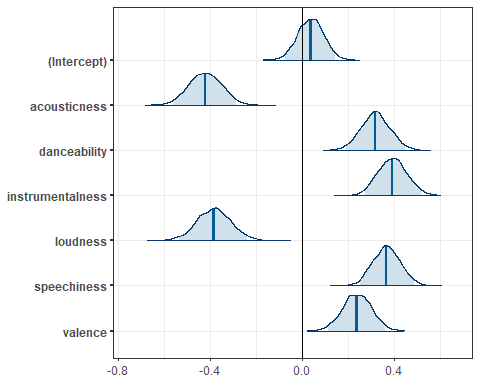
|  |  |  |
| --- | --- | --- |
| term | estimate | std.error |
| (Intercept) | 0.044 | 0.055 |
| acousticness | -0.423 | 0.073 |
| danceability | 0.311 | 0.063 |
| duration\_ms | 0.218 | 0.063 |
| instrumentalness | 0.360 | 0.066 |
| loudness | -0.360 | 0.077 |
| speechiness | 0.389 | 0.060 |
| valence | 0.270 | 0.063 |

### Model with Horseshoe Prior

## Sampling priors, please wait...

## Bayes Factor (Savage-Dickey density ratio)  
##   
## Parameter | BF  
## -------------------------  
## (Intercept) | 0.023  
## acousticness | > 1000  
## danceability | > 1000  
## instrumentalness | > 1000  
## loudness | > 1000  
## speechiness | > 1000  
## valence | > 1000  
##   
## \* Evidence Against The Null: 0

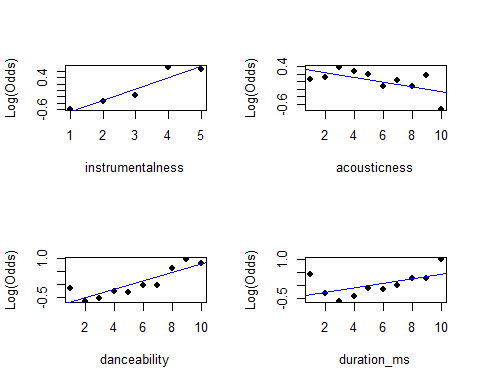
## Call:  
## BF.glm(x = hs\_model)  
##   
## Bayesian hypothesis test  
## Type: exploratory  
## Object: stanreg  
## Parameter: general parameters  
## Method: adjusted fractional Bayes factors using Gaussian approximations  
##   
## Posterior probabilities:  
## Pr(=0) Pr(<0) Pr(>0)  
## (Intercept) 0.940 0.015 0.045  
## acousticness 0.000 1.000 0.000  
## danceability 0.000 0.000 1.000  
## instrumentalness 0.000 0.000 1.000  
## loudness 0.000 1.000 0.000  
## speechiness 0.000 0.000 1.000  
## valence 0.016 0.000 0.984



##   
## Model Info:  
## function: stan\_glm  
## family: binomial [logit]  
## formula: target ~ acousticness + danceability + instrumentalness + loudness +   
## speechiness + valence  
## algorithm: sampling  
## sample: 4000 (posterior sample size)  
## priors: see help('prior\_summary')  
## observations: 1513  
## predictors: 7  
##   
## Estimates:  
## mean sd 10% 50% 90%  
## (Intercept) 0.0 0.1 0.0 0.0 0.1   
## acousticness -0.4 0.1 -0.5 -0.4 -0.3   
## danceability 0.3 0.1 0.2 0.3 0.4   
## instrumentalness 0.4 0.1 0.3 0.4 0.5   
## loudness -0.4 0.1 -0.5 -0.4 -0.3   
## speechiness 0.4 0.1 0.3 0.4 0.4   
## valence 0.2 0.1 0.2 0.2 0.3   
##   
## Fit Diagnostics:  
## mean sd 10% 50% 90%  
## mean\_PPD 0.5 0.0 0.5 0.5 0.5   
##   
## The mean\_ppd is the sample average posterior predictive distribution of the outcome variable (for details see help('summary.stanreg')).  
##   
## MCMC diagnostics  
## mcse Rhat n\_eff  
## (Intercept) 0.0 1.0 6021   
## acousticness 0.0 1.0 3122   
## danceability 0.0 1.0 3627   
## instrumentalness 0.0 1.0 3687   
## loudness 0.0 1.0 3488   
## speechiness 0.0 1.0 3465   
## valence 0.0 1.0 3279   
## mean\_PPD 0.0 1.0 5082   
## log-posterior 0.1 1.0 1303   
##   
## For each parameter, mcse is Monte Carlo standard error, n\_eff is a crude measure of effective sample size, and Rhat is the potential scale reduction factor on split chains (at convergence Rhat=1).

|  |  |  |
| --- | --- | --- |
| term | estimate | std.error |
| (Intercept) | 0.037 | 0.057 |
| acousticness | -0.422 | 0.072 |
| danceability | 0.319 | 0.063 |
| instrumentalness | 0.392 | 0.063 |
| loudness | -0.384 | 0.075 |
| speechiness | 0.367 | 0.061 |
| valence | 0.239 | 0.063 |

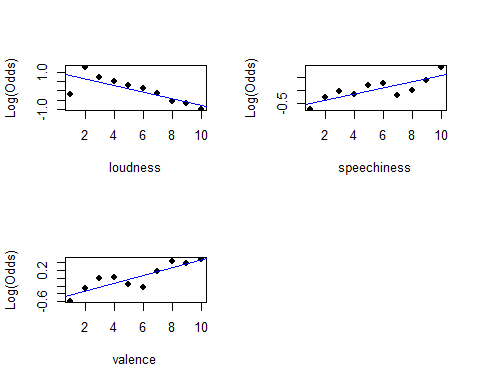
## Model Conditions



## Registered S3 methods overwritten by 'car':  
## method from  
## influence.merMod lme4  
## cooks.distance.influence.merMod lme4  
## dfbeta.influence.merMod lme4  
## dfbetas.influence.merMod lme4

## # A tibble: 7 x 2  
## Variables VIF  
## <chr> <dbl>  
## 1 acousticness 1.57  
## 2 danceability 1.25  
## 3 duration\_ms 1.17  
## 4 instrumentalness 1.29  
## 5 loudness 1.83  
## 6 speechiness 1.06  
## 7 valence 1.30

## # A tibble: 6 x 2  
## Variables VIF  
## <chr> <dbl>  
## 1 acousticness 1.63  
## 2 danceability 1.23  
## 3 instrumentalness 1.25  
## 4 loudness 1.86  
## 5 speechiness 1.05  
## 6 valence 1.29



Now that our models are created, we check our model conditions for logistic regression. The first condition is linearity. As is shown by the graphs above, each of the continuous predictors in our models appear to have a more or less linear relationship with the log odds of the person liking a song. As a result, the linearity condition is satisfied.

The next condition is the randomness condition. Because the samples are not stated to be truly randomly taken we must consider whether or not the observations differ systemically from our population of interest. The population of interest is songs that the person who made the data set listens to and rates as a like or a dislike, and so there is no reason to believe that the sample is not representative of the population. The randomness condition is satisfied.

The next condition is the independence condition. We have no reason to believe that each song is not independent from one another, and so the independence condition is satisfied.

Lastly, we check multicollinearity by checking the variance inflation factors of each of the variables. As seen in the output for both cauchy and horseshoe, since all VIF values are under 10, multicollinearity is not a problem in our model.

## Model Comparison and Results

##   
## Computed from 4000 by 1513 log-likelihood matrix  
##   
## Estimate SE  
## elpd\_loo -944.9 14.2  
## p\_loo 8.7 0.4  
## looic 1889.7 28.4  
## ------  
## Monte Carlo SE of elpd\_loo is 0.0.  
##   
## All Pareto k estimates are good (k < 0.5).  
## See help('pareto-k-diagnostic') for details.

##   
## Computed from 4000 by 1513 log-likelihood matrix  
##   
## Estimate SE  
## elpd\_loo -950.1 13.4  
## p\_loo 7.6 0.3  
## looic 1900.1 26.8  
## ------  
## Monte Carlo SE of elpd\_loo is 0.0.  
##   
## All Pareto k estimates are good (k < 0.5).  
## See help('pareto-k-diagnostic') for details.

##   
## Computed from 4000 by 1513 log-likelihood matrix  
##   
## Estimate SE  
## elpd\_loo -1049.6 0.5  
## p\_loo 1.0 0.0  
## looic 2099.2 0.9  
## ------  
## Monte Carlo SE of elpd\_loo is 0.0.  
##   
## All Pareto k estimates are good (k < 0.5).  
## See help('pareto-k-diagnostic') for details.

To compare the predictive performance of our three models, we use leave-one-out cross-validation (LOO). In particular, we use the LOO Information Criterion (LOOIC) as a metric by which to measure the predictive performance of our models. LOOIC measures the predictive accuracy of a given model by leaving a single sample out of the training set and training the model on the all other samples in the dataset. It then tests its accuracy on this single left-out sample and repeats this process for all other samples in the training set. LOOIC reports a metric based on the accuracy of these LOO predictions, where a lower LOOIC score corresponds to a better model (Gabry & Goodrich). The LOOIC values for all three models are shown below. Of the three models, we see that the cauchy prior model performed the best.

|  |  |  |
| --- | --- | --- |
| Cauchy | Horseshoe | Null |
| 1889.7 | 1900.13 | 2099.25 |

With our chosen model, we now want to be able to use it on song data we may find in the wild. However, our current model was fit on a normalized dataset, so its parameters are only trained to classify songs by characteristics with mean 0 and standard deviation 1. To retrieve the parameters for a model that works on unnormalized data () instead of normalized data , we adjust our coefficients as follows:

With our unnormalized model, we first check to see how well it predicts the ‘target’ value of the samples on which it was trained. Using a probability threshold of 0.5, we find that our cauchy prior predictor categorizes 0.664 of the songs in the training set. For comparison, the horseshoe prior model has a prediction accuracy of 0.662, and the “null” model has a prediction accuracy of 0.506 on the training set. Of course, here we are checking performance on songs which the models have already trained on. What we really want to measure is the ability of our models to make predictions on songs they have not yet seen. For this, we check the accuracy of our models on the test set. Since our models have not seen this data yet, we would expect their performance on this dataset to be slightly worse than on the training set. This is exactly what we see for the cauchy and horseshoe prior models. Again, using a probability threshold of 0.5, we find that our cauchy prior predictor has accuracy 0.6329365 on songs in the test set. For comparison, the horseshoe prior model has prediction accuracy hs\_test\_acc, and the “null” model has prediction accuracy of one\_test\_acc on the test set. Thus, we see that under this metric, the cauchy prior predictor again performs the best.

We also find that all of our models seem to perform better on negative examples than on positive one. The false positive rate for our cauchy prior model was model\_test\_fp, while its false negative rate was model\_test\_fn. The values for all models are given in the table below. While the cauchy prior model still performed best on both metrics, this trend exists for all three models, which seems to imply that songs this person dislikes are easier to categorize based on song attributes alone than songs he likes.

|  |  |  |  |
| --- | --- | --- | --- |
| Class Error | Cauchy | Horseshoe | Null |
| False Positive | 0.3636 | 0.3815 | 0.496 |
| False Negative | 0.3705 | 0.3922 | Undefined |

## Conclusion

Our final model with a weakly informative cauchy prior had a predictive accuracy of 0.6329365 on the testing data. This accuracy is noticeably better than our null model that guesses 1 every time and ends up with an average prediction accuracy of a little over 50%, a result of the data being comprised of a little over 50% liked songs.

This prediction/recommendation accuracy would be acceptable to us as Spotify users. From anecdotal evidence, while Spotify does recommend plenty of songs that we enjoy, it also recommends many songs that we do not like as well. From being active Spotify users, we would estimate the accuracy of Spotify’s recommendations to be in the same ballpark as our predictive accuracy. Spotify also has the added bonus of having extra variables to create their recommendation that we were not able to use. For example, Spotify is able to take into account similar user’s activities as well as information about artists and genres that we were not able to use. However, we do have the advantage of explicitly knowing whether or not a song is liked or disliked, while Spotify must infer that information for each user. Because our model was built given this information, it does not have the capability to be used as a replacement to Spotify’s algorithm because users will not explicitly record whether or not they like each song.

For further exploration on this topic, if we had access to data regarding listening times, whether or not a song is on a user’s playlist, and other such data that might clue us in on if a user enjoys a song, we could create an algorithm with this data to figure out if a user likes a song, and then create recommendations with our current logistic models using the values of ‘target’ that we predicted. This would probably drop our accuracy rate, but it would then have the potential for real world use within the Spotify app that Spotify’s algorithms have. Until then, our algorithm can be used for people like the Kaggle user who compiled information about songs he or she did or didn’t like.