Lesson 24 Spotify Solution

LTC J.K. Starling

Last compiled on 31 March, 2022

Part I: Background

Spotify recommends songs based on user listening history. The algorithms use the characteristics of the music a user is known to like to figure out the music the user may like but has not discovered yet.

In this lab you will use logistic regression to build a model that predicts the probability a user likes a song based on its characteristics. The data is from the Spotify Song Attributes data set which contains the song preferences of a single Spotify user. The data set contains the characteristics of 2017 songs and whether the Spotify user liked the song (1-like , 0-dislike).

Some of the available attributes are:

- Danceability: How suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- Acousticness: A measure from 0.0 to 1.0 of whether the track is acoustic.
- Energy: A measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
- Instrumentalness: Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.
- Key: Denotes the major or minor scale in which a piece of music operates, and thus, the notes that belong in it.
- Liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
- Loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track. Values typical range between -60 and 0 db.
- Mode: Indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived.
- Speechiness: Detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value.
- Tempo: The overall estimated pace of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

(1) Use the definitions in the documentation to classify the following variables: instrumentalness, key, danceability, mode, and tempo?

Instrumentalness: Quantitative Key: Categorical Danceability: Quantitative Mode: Categorical Tempo: Quantitative

Part II: Exploratory Data Analysis

Before you begin your analysis, be sure to change the categorical variables you will use into factors.

(2) Create a) a boxplot to look at the distribution of 'danceability' based on whether the Spotify user liked the song and b) a bar graph to look at the distribution of 'key' based on whether the Spotify user liked the song.

```
# class(music$like)
music <- music %>% mutate(like = as.factor(like), key = as.factor(key), mode = as.factor(mode))
music %>% ggplot(aes(x = like, y = danceability)) +
  geom_boxplot() +
  labs(Title = "Danceability as an indicator of like", x = "Like", y = "Danceability")
```

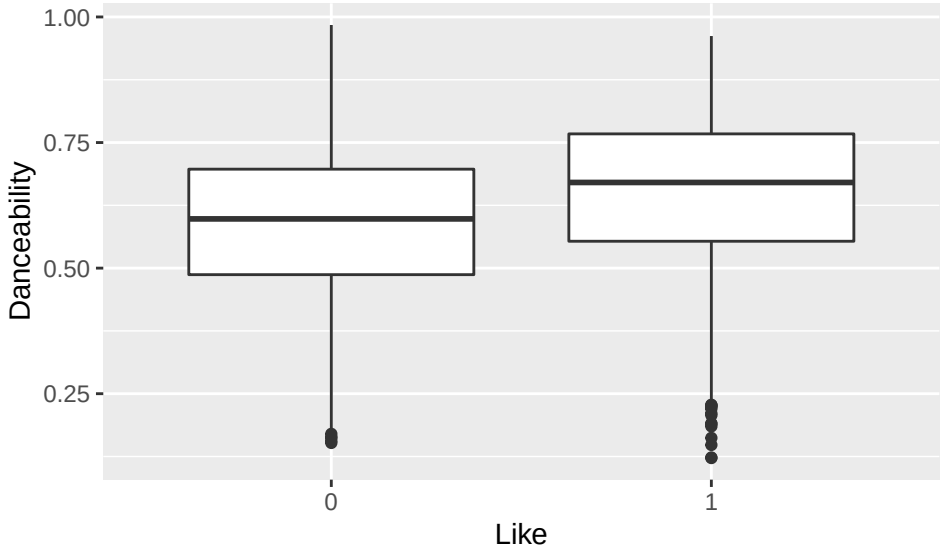

Figure 1: Danceability by Like

```
music %>% ggplot(aes(x = key)) +
  geom_bar() +
  facet_wrap(~like) +
  labs(Title = "Key as an indicator of like", x = "key", y = "count")
```

(3) Use no more than five sentences to summarize your observations from both plots.

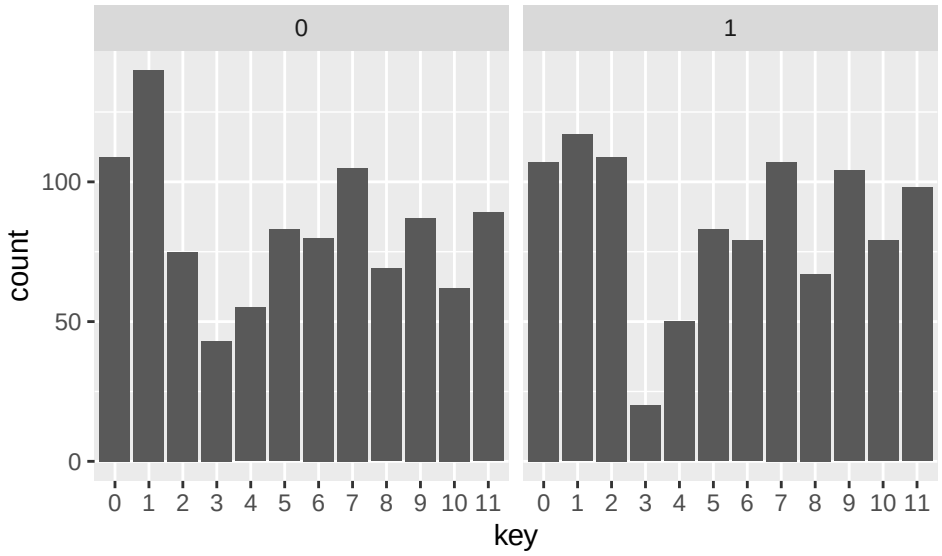

Figure 2: Like vs. Key

Looking at the box plot, likes have a higher median danceability rating when compared to songs that are not liked. Furthermore, seem to have more outliers than on-liked songs.

Next, looking at the bar graph, the similar distributions by key are similar for liked and non-liked songs.

Part III: Logistic Regression Model

(4) Write the general form of the logistic regression model that includes every explanatory variable in question 1. (Use π as the probability the Spotify user will like the song.)

(5) Fit the model in question 4.

```
model1 <- music %>%
  glm(like~instrumentalness+key+danceability+mode+tempo,data=.,family='binomial')

summary(model1)
```

```
##
## Call:
## glm(formula = like ~ instrumentalness + key + danceability +
##     mode + tempo, family = "binomial", data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9301  -1.1126   0.5337   1.1096   1.9560
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.352423   0.355260  -6.622 3.55e-11 ***
```

```
## instrumentalness  1.482539   0.185974   7.972 1.56e-15 ***
## key1             -0.203065   0.192070  -1.057   0.2904
## key2              0.497976   0.209769   2.374   0.0176 *
## key3             -0.780343   0.315581  -2.473   0.0134 *
## key4             -0.091629   0.248624  -0.369   0.7125
## key5             -0.065271   0.217086  -0.301   0.7637
## key6             -0.049381   0.218627  -0.226   0.8213
## key7              0.021273   0.200833   0.106   0.9156
## key8             -0.135769   0.227559  -0.597   0.5508
## key9              0.182064   0.207586   0.877   0.3805
## key10             0.140836   0.228819   0.615   0.5382
## key11             0.058313   0.211053   0.276   0.7823
## danceability      2.814838   0.309377   9.098  < 2e-16 ***
## mode1            -0.256481   0.099640  -2.574   0.0101 *
## tempo             0.004884   0.001770   2.759   0.0058 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2795.9  on 2016  degrees of freedom
## Residual deviance: 2624.9  on 2001  degrees of freedom
## AIC: 2656.9
##
## Number of Fisher Scoring iterations: 4
```

```
coef(model1)
```

```
##      (Intercept) instrumentalness             key1             key2
##     -2.352422735      1.482538738     -0.203064538      0.497975958
##             key3             key4             key5             key6
##     -0.780343119     -0.091629452     -0.065271372     -0.049381166
##             key7             key8             key9            key10
##      0.021272920     -0.135769224      0.182063620      0.140836149
##            key11     danceability            mode1            tempo
##      0.058313262      2.814838453     -0.256481070      0.004883717
```

```
confint(model1)
```

```
##                       2.5 %       97.5 %
## (Intercept)      -3.05369208 -1.660405445
## instrumentalness  1.12272672  1.852375312
## key1             -0.58018218  0.173230412
## key2              0.08813971  0.911094808
## key3             -1.41465339 -0.172957606
## key4             -0.58059234  0.395370062
## key5             -0.49133303  0.360352133
## key6             -0.47842772  0.379324808
## key7             -0.37255827  0.415263720
## key8             -0.58260810  0.310329776
## key9             -0.22448099  0.589866364
## key10            -0.30713686  0.590723026
## key11            -0.35537531  0.472569555
```

```
## danceability      2.21338199  3.426640349
## mode1            -0.45201967 -0.061327667
## tempo             0.00142075  0.008362802
```

- Interpret the coefficient for danceability and comment on it's statistical significance at the 0.05 level.

- Report and interpret the odds ratio and the associated 95% confidence interval.

(6) Explain what it means for there to be an interaction between mode and tempo. Use a model to evaluate whether it is reasonable to believe that there is an interaction between ode and tempo. Be sure to provide supporting evidence for your response.

```
# Interaction model
int.glm <- music %>% glm(like~mode*tempo, data=., family = 'binomial')
summary(int.glm)
```

```
##
## Call:
## glm(formula = like ~ mode * tempo, family = "binomial", data = .)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.295  -1.165   1.069   1.180   1.294
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.0967144  0.3416803   0.283    0.777
## mode1       -0.5733726  0.4319200  -1.327    0.184
## tempo        0.0008788  0.0027036   0.325    0.745
```

```
## mode1:tempo  0.0023135  0.0034496   0.671    0.502
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2795.9  on 2016  degrees of freedom
## Residual deviance: 2783.0  on 2013  degrees of freedom
## AIC: 2791
##
## Number of Fisher Scoring iterations: 3
```

```
##Interpretation

##The interaction term accounts for changes in two explanatory variables.
##The effect of one variable (mode or tempo) depends on the values of the other (tempo or mode)

##The p value suggests that there is no statistically significant interaction between these variables
##pval > 0.05
##z values are relatively low (zval<3)
##Additionally, the AIC and null deviance are higher than the first model suggesting that the interacti
```

Part IV: Model Assessment

(7) Recall that the primary goal of this analysis is to predict the probability this Spotify user will like a song. Regardless of your responses in previous questions, let us pretend that both the models in question 4 and 6 are valid for inference. What metric could you use to select the better model for prediction?

Higher R^2 value and lower standard deviation. Specifically, we looked at the AIC value, as it accounts for the increase in the R^2 value due to the inclusion of additional variables regardless of whether or not they are helpful. We also referenced the deviance to compare the models (lower being better).

Note to students: The null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean) where as the residual deviance shows how well the inclusion of independent variables predicts the response variable.

(8) Write the equation to calculate the predicted probability the Spotify user likes a particular song based on the preferred model.

$\pi = \frac{e^{\beta_0 + beta_1(instrumentalness) + beta_2(key) + beta_3(dance) + beta_4(mode) + beta_5(tempo)}}{1 + e^{\beta_0 + beta_1(instrumentalness) + beta_2(key) + beta_3(dance) + beta_4(mode) + beta_5(tempo)}}$

pi = odds/1 + odds

(9) Use a cutoff of 0.5 to calculate the correct classification rate for the preferred model.

```
# Calculating correct classification rate
cutoff = 0.5

model1.pred <- ifelse(model1$fitted.values >= cutoff, "like.p", "dislike.p")
A <- table(music$like, model1.pred)
A.TP <- tryCatch(A[1, "like.p"], error = function(ee) 0)
A.TN <- tryCatch(A[0, "dislike.p"], error = function(ee) 0)
model1.rate <- (635+623)/2017

int.glm.pred <- ifelse(int.glm$fitted.values >= cutoff, "like.p", "dislike.p")
B <- table(music$like, int.glm.pred)
B.TP <- tryCatch(B[1, "like.p"], error = function(ee) 0)
B.TN <- tryCatch(B[0, "dislike.p"], error = function(ee) 0)
int.glm.rate <- (633+617) / 2017
```

(10) How well does the model distinguish between two songs: "Sign of the Times" by Harry Styles and "Hotline Bling" by Drake. Briefly explain using supporting evidence. (Insert the preferred model in the code below to calculate the probability the user likes the two songs).

```
# Read data for "Sign of the Times" by Harry Styles and "Hotline Bling" by Drake
test_songs <- read_csv("https://raw.githubusercontent.com/jkstarling/MA376/main/test_songs.csv")
test_songs <- test_songs %>% mutate(mode = as.factor(mode), key = as.factor(key))

# Predicted probability based on preferred model
predict.glm(object = model1, newdata = test_songs, type = "response")
```

```
##         1         2
## 0.3461945 0.7447236
```