Nirali Bandaru Rohan Gangisetty

Plan:

Models we know:

1. Linear Regression
2. Multiple linear regression
3. Logistic regression (classification)
4. Non-linear models
    a. Polynomial Regression
    b. Smoothing spline
    c. Regression spline
    d. Piecewise polynomial regression
    e. Generalized Additive models

How is our response variable measured?

- Quantitative: popularity is ranged from 0 - 100

Observations made in EDA:

1. Danceability is all over the place
2. The other two are skewed

Must do:

1. Observe other features
2. Cross-correlation between predictors - cor() function
3. Review non-linear models
    a. Polynomial Regression with regression splines
4. Review each feature vs. response variable plot to observe trend, to see what function fits best to the visual.
5. Resampling methods:
    a. K-fold
    b. LOOCV
6. Possible methods:
    a. LOOCV - not possible because too large of a dataset, resulting in high computational complexity
    b. K-fold - could do
        i. 5 groups
        ii. 10 groups
        iii. 8 decades ~ 8 groups
    c. Normal split between training and testing
        i. 70-30%
        ii. 60-40%
        iii. 50-50%
        Test to see best combo
7. Fit model
8. Make predictions
    a. 3 cases of interest (?????)

Nirali Bandaru Rohan Gangisetty

GROUP ASSIGNMENT 3 - SPOTIFY 2

You summary should include the following information:

- Justify your model choice based on how your response is measured and any observations you may have made in your EDA.
- Report the model's test error rate using one of the techniques we discussed in lecture. Justify your choice.
- Based on the estimated test error rate, discuss how well the model fits the data.
- Use the model to make predictions for at least three cases of interest.

EDA Commentary

Response Variable - Popularity

The response variable, popularity, is distributed nearly evenly across the popularity index, but there are a maximum number of songs that are unpopular, or with a popularity index of zero. This is illustrated by the frequency distribution in the histogram plot of the response variable.

Predictors

Predictor	**Observation**
Danceability	No apparent pattern
Loudness	Distribution skewed right, nearly all the songs are on the top half of the loudness index
Speechiness	Distribution skewed left, most points on the bottom half of the speechiness index
Liveness	Distribution suggests most songs are popular if the liveness index is fairly low. Top half of the distribution still has a significant number of data but is less dense when compared with the bottom half.
Acousticness	No apparent pattern
Key	Similarly distributed, no observed extremities
Instrumentalness	Distribution is dense towards the ends of the instrumentalness index, less dense in the middle. Shape could be distributed as a U.
Duration	Slightly peaks at the middle, where duration is approximately 3 min
Tempo	Is condensed slightly inward. Extreme tempos at 0-40 and 225-250 have barely any data points. Otherwise, no apparent pattern is observed between tempo and popularity.

One change was made since exploratory data analysis was performed on the data set: The predictor "mode" was removed due to its lack of correlation with the response variable.

Model Chosen: Polynomial Regression

Discussion

The response variable in this study is quantitative as are many of the predictors. Upon careful examination and analysis of each predictor vs. response variable plot, it was concluded that the relationships look mostly non-linear. Classification is used for categorical variables and would have proven more appropriate had the output index been divided into intervals, where each interval would be a category. However, since the response variable was considered purely quantitative, it was more appropriate to choose a regression model. Among the pool of regression models, polynomial regression was chosen for its ability to train and predict nonlinear relationships.

Model Parameters and Results

Resampling and Cross-Validation

Cross-validation was done using the k-fold method due to the vast number of test points in the data set. Number of folds was chosen at K = 10 and using method = “glm.”
After doing some research, it was found that K = 5 or K = 10 was optimal for large datasets. Since there are a total of nine predictors in the data set, K = 10 was chosen as a reference value.
The degree of polynomial was chosen to be at 4. Changing the degree value reflected very little change in the r-squared value. Hence, d = 4 was kept consistent.

R-squared value
R-squared value : 0.4203
RMSE: 16.47

Analysis of Results
The r-squared value is lower than 0.5, which shows that the model is not a good fit for the data.

Predictions

RESOURCES

Ggplot help:
http://www.sthda.com/english/wiki/paired-samples-t-test-in-r
https://community.rstudio.com/t/boxplot-error-must-request-at-least-one-color-from-a-hue-palette/33869/2
http://environmentalcomputing.net/plotting-with-ggplot-adding-titles-and-axis-names/#:~:text=To%20alter%20the%20labels%20on,line%20of%20basic%20ggplot%20code.&text=Note%3A%20You%20can%20also%20use,which%20is%20equivalent%20to%20ggtitle%20.

Musical keys for key plot legend:
http://musictheoryfundamentals.com/MusicTheory/keySignatures.php
https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

Pitch class:
https://en.wikipedia.org/wiki/Pitch_class