4_Detailed_Report

Hyungjae Kim

2022-08-12

Dataset

Raw Dataset

Removed `track_hist`, `url`, `artists_ids`, `track_id`, `chart_start`, `chart_end`, and `new_id`.

Classified Dataset

Classified songs as hit songs by using a `hitness` threshold of 250, converted duration from milliseconds to minutes, and log scaled popularity and tempo. This dataset was used for all the models.

Log Scaling Popularity

Log Scaling Tempo

Number of Rows / Feature Vectors

Summary of Raw Data (Continuous Hitness Values and Unscaled)

```
    hitness        danceability        energy            key        
 Min.   :  20.0   Min.   :0.0671   Min.   :0.0326   Min.   : 0.000  
 1st Qu.:  40.0   1st Qu.:0.5290   1st Qu.:0.5700   1st Qu.: 2.000  
 Median : 220.0   Median :0.6300   Median :0.7090   Median : 5.000  
 Mean   : 419.5   Mean   :0.6251   Mean   :0.6822   Mean   : 5.208  
 3rd Qu.: 480.0   3rd Qu.:0.7290   3rd Qu.:0.8180   3rd Qu.: 8.000  
 Max.   :4250.0   Max.   :0.9860   Max.   :0.9960   Max.   :11.000  
    loudness            mode         speechiness      acousticness      
 Min.   :-29.224   Min.   :0.0000   Min.   :0.0225   Min.   :0.0000033  
 1st Qu.: -7.054   1st Qu.:0.0000   1st Qu.:0.0368   1st Qu.:0.0183000  
 Median : -5.616   Median :1.0000   Median :0.0563   Median :0.0736000  
 Mean   : -5.964   Mean   :0.6715   Mean   :0.1062   Mean   :0.1718459  
 3rd Qu.: -4.403   3rd Qu.:1.0000   3rd Qu.:0.1280   3rd Qu.:0.2410000  
 Max.   :  0.175   Max.   :1.0000   Max.   :0.9510   Max.   :0.9930000  
 instrumentalness      liveness         valence           tempo       
 Min.   :0.0000000   Min.   :0.0193   Min.   :0.0349   Min.   : 48.72  
 1st Qu.:0.0000000   1st Qu.:0.0979   1st Qu.:0.3190   1st Qu.:100.01  
 Median :0.0000000   Median :0.1310   Median :0.4780   Median :124.08  
 Mean   :0.0112172   Mean   :0.1892   Mean   :0.4861   Mean   :123.97  
 3rd Qu.:0.0000131   3rd Qu.:0.2460   3rd Qu.:0.6510   3rd Qu.:143.89  
 Max.   :0.9550000   Max.   :0.9790   Max.   :0.9760   Max.   :213.74  
```

```
  duration_ms     time_signature    popularity       sentiment
 Min.   : 46253   Min.   :1.000   Min.   :     1   Min.   :-1.0000
 1st Qu.:197759   1st Qu.:4.000   1st Qu.:  2938   1st Qu.:-0.7613
 Median :219840   Median :4.000   Median :  7716   Median : 0.9636
 Mean   :224905   Mean   :3.973   Mean   : 13929   Mean   : 0.3771
 3rd Qu.:245867   3rd Qu.:4.000   3rd Qu.: 16468   3rd Qu.: 0.9949
 Max.   :688453   Max.   :5.000   Max.   :115282   Max.   : 1.0000
```

Summary of Classified and Cleaned Data (Factorized and Scaled)

Linear Model (OLS)

Accuracy

Summary (Coefficients, F-Statistics, P-Value)

Anova Table

Plots (Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage)

Logistic Regression

Accuracy

Summary (Coefficients, F-Statistics, P-Value)

Anova Table

Plots (Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage)

Logistic LASSO Regression

Accuracy

LASSO Log Selection (1SE)

Summary (Coefficients, F-Statistics, P-Value)

Anova Table

Plots (Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage)