Herbert González Rionda
Professor Evgeniya Reshetnyak
APANPS 5100-002
29 November 2022

# Predictive Analysis Competition: Report

For the Predictive Analysis Competition on "The Perfect Tune", I decided to focus on developing a linear model following the insights of author Josh Zumbrun at The Wall Street Journal in his article titled "When It Comes to Data, Sometimes Less Is More". The title of the article says it all, and in the case of the model for my best submission on Kaggle, it holds true in terms of its simplicity.

My best submission throughout the competition held on Kaggle, submission 6, the RMSE was that of **15.39157**. The linear model for submission 6 can be seen below:

*model6 =*
*lm(rating~tempo+loudness+energy+liveness+loudness+track_duration+danceability+instrumentalness+valence+tempo+time_signature+pop+dance_pop+pop_rock+rnb+alternative_hip_hop+urban_contemporary+prog_electro_house, analysis)*

Model 6 was directly developed from linear model 5, which can be seen below and holds an RMSE of 15.60746:

*model5 =*
*lm(rating~tempo+loudness+energy+liveness+loudness+track_duration+danceability+instrumentalness+valence+tempo+time_signature, analysis)*

The difference between the two models is the addition of various variables in the context of music family that hold important significant codes on the summary of a model holding all of the variables regarding music type. The model use to test the significant codes can be seen below:

*modeltrial =*
*lm(rating~tempo+loudness+energy+liveness+loudness+track_duration+danceability+instrumentalness+valence+tempo+time_signature+pop+dance_pop+house+teen_pop+electro_house+edm+pop_rap+pop_rock+rnb+alternative_hip_hop+urban_contemporary+prog_electro_house+indie_rnb, analysis)*

The summary of this model shows the following coefficients with the significant codes, with the most important ones being used in model 6:

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 1.923e+01 | 1.649e+00 | 11.660 | < 2e-16 | *** |
| tempo | 1.935e-02 | 4.125e-03 | 4.691 | 2.74e-06 | *** |
| loudness | 6.027e-01 | 4.775e-02 | 12.621 | < 2e-16 | *** |
| energy | -2.002e+00 | 9.289e-01 | -2.155 | 0.03114 | * |
| liveness | -3.505e+00 | 7.085e-01 | -4.947 | 7.59e-07 | *** |
| track_duration | 2.540e-05 | 1.736e-06 | 14.630 | < 2e-16 | *** |
| danceability | 1.456e+01 | 8.885e-01 | 16.382 | < 2e-16 | *** |
| instrumentalness | -6.257e+00 | 8.238e-01 | -7.596 | 3.20e-14 | *** |
| valence | -7.769e+00 | 6.078e-01 | -12.784 | < 2e-16 | *** |

```
time_signature       2.727e+00  3.639e-01   7.496 6.87e-14 ***
pop                 2.092e+00  2.605e-01   8.029 1.04e-15 ***
dance_pop            5.254e+00  4.645e-01  11.310  < 2e-16 ***
house                8.696e-01  9.603e-01   0.906  0.36521
teen_pop             4.304e-01  5.496e-01   0.783  0.43362
electro_house       -3.793e+00  2.755e+00  -1.377  0.16861
edm                 -6.252e-02  1.427e+00  -0.044  0.96505
pop_rap              3.815e-01  4.475e-01   0.852  0.39395
pop_rock             6.158e+00  5.555e-01  11.086  < 2e-16 ***
rnb                 3.607e+00  6.779e-01   5.321 1.05e-07 ***
alternative_hip_hop -5.543e+00  1.848e+00  -2.999  0.00271 **
urban_contemporary  -2.587e+00  6.636e-01  -3.899 9.71e-05 ***
prog_electro_house   1.123e+01  4.570e+00   2.457  0.01402 *
indie_rnb           -3.725e+00  4.192e+00  -0.889  0.37415
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The final variables used in model 6 all had some level of significance, as evidenced by the summary of the model below:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.917e+01  1.643e+00  11.665  < 2e-16 ***
tempo               1.969e-02  4.113e-03   4.786 1.71e-06 ***
loudness            6.094e-01  4.701e-02  12.964  < 2e-16 ***
energy             -2.024e+00  9.253e-01  -2.188 0.028703 *
liveness           -3.485e+00  7.070e-01  -4.929 8.33e-07 ***
track_duration      2.543e-05  1.734e-06  14.668  < 2e-16 ***
danceability        1.482e+01  8.532e-01  17.365  < 2e-16 ***
instrumentalness   -6.238e+00  8.235e-01  -7.575 3.76e-14 ***
valence            -7.900e+00  5.937e-01 -13.307  < 2e-16 ***
time_signature      2.732e+00  3.637e-01   7.511 6.12e-14 ***
pop                2.168e+00  2.483e-01   8.731  < 2e-16 ***
dance_pop           5.419e+00  4.073e-01  13.303  < 2e-16 ***
pop_rock            6.093e+00  5.485e-01  11.108  < 2e-16 ***
rnb                3.477e+00  6.617e-01   5.255 1.50e-07 ***
alternative_hip_hop -5.595e+00  1.848e+00  -3.028 0.002463 **
urban_contemporary  -2.498e+00  6.585e-01  -3.794 0.000149 ***
prog_electro_house   8.122e+00  3.874e+00   2.097 0.036039 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In an effort to further fine tune the next model after constructing model 6, I decided that I would eliminate the variables that had a significance level of 0.01; these being energy and prog_electro_house. The intention behind reducing the number of variables using the previous method discussed was to lower the RMSE. However, eliminating those two variables in model 7 slightly increased the RMSE to 15.39708; meaning that my methodology was losing precision.

The failed missteps along the way included either using too little or too much data in the model. Model 6 was the best at creating a balanced model that would correctly predict the rating of a song based on the interaction between the variables it analyzed; it was in m opinion, not too overpopulated and not too underpopulated.

In the spirit of the words of Josh Zumbrun which hold strength in the methodology I used to develop the models for this competition, if I had to do something differently I would dive deeper into analyzing the genres and singers while still using an lm() function. It is clear that the linear model holds strength for PAC on “The Perfect Tune”, but additional analytical efforts and methodology could be applied to the genre and artist variables to produce extra features. I would transform the data in the “genre” and “performer” columns into their own columns in order to be able to use them in my analysis. By incorporating the data in “genre” and “performer” as variables that can be used in a linear model, I would be able to further apply my strategy of analyzing these variables through their significant codes and better my current linear model in an effort to find the perfect tune.

Works Cited

Zumbrun, Josh. “When It Comes to Data, Sometimes Less Is More.” *The Wall Street Journal*, Dow Jones & Company, 4 Nov. 2022, https://www.wsj.com/articles/when-it-comes-to-data-sometimes-less-is-more-11667554203.