Popularity Index Analysis

Marc Schmieder

2/28/2022

1. Exploration

1.1 Dataset attributes

This document summarizes a first draft from the Spotify Popularity Index (PI) analysis. In the current state, 331 data points from the community were provided, whereas each data point equals one song at a given time. The data points were anonymized by creating an artist/song ID. Overall, 31 variables (columns) were provided or engineered through existing data. The column names are

```
##  [1] "ArtistSongId"
##  [2] "PopularityIndex"
##  [3] "Timestamp"
##  [4] "DaysSinceRelease"
##  [5] "ReleaseDate"
##  [6] "StreamsLast28Days"
##  [7] "ListenersLast28Days"
##  [8] "SavesLast28Days"
##  [9] "StreamsAllTime"
## [10] "ListenersAllTime"
## [11] "NumberOfBlogsThatCoveredTheSong"
## [12] "NumberOfPlaylistsAllTime"
## [13] "EmailAddress"
## [14] "CurrentSpotifyFollowers"
## [15] "PopularityIndexSource"
## [16] "StreamsLast7Days"
## [17] "ListenersLast7Days"
## [18] "SavesLast7Days"
## [19] "DiscoverWeeklyStreamsLast28Days"
## [20] "DiscoverWeeklyStreamsLast7Days"
## [21] "ReleaseRadarStreamsLast28Days"
## [22] "ReleaseRadarStreamsLast7Days"
## [23] "NumberOfBlogsThatCoveredTheSong_num"
## [24] "StreamsLast28Days_PerListener"
## [25] "SavesLast28Days_PerListener"
## [26] "StreamsAllTime_PerListener"
## [27] "NumberOfPlaylistsAllTime_PerListener"
## [28] "StreamsLast7Days_PerListener"
## [29] "SavesLast7Days_PerListener"
## [30] "PopularityIndexSource_Bin"
## [31] "ReleasePassed21Days"
```

As to be seen later in the analysis, the variables

```
c("StreamsLast28Days", "ListenersLast28Days",
  "StreamsLast7Days", "ListenersLast7Days")
```

```
## [1] "StreamsLast28Days"   "ListenersLast28Days" "StreamsLast7Days"
## [4] "ListenersLast7Days"
```

are the most important ones for the later conducted model, predicting the Popularity Index

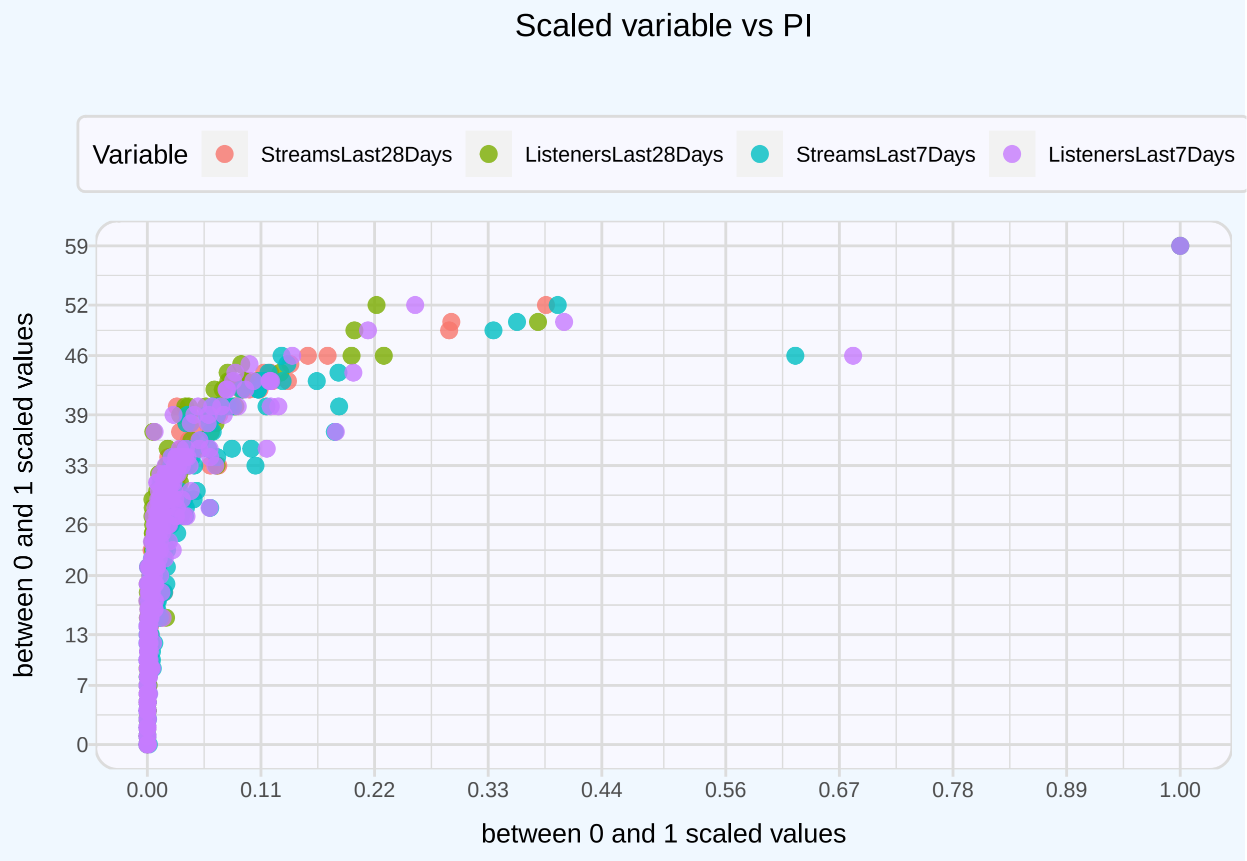
1.2 Mean values for different PI

As certain PI numbers are criterias for getting into algorithmic playlists, the following table shows the mean values in the data for PI indices of 20-22, 30-32 and 40-42. With that information, an artist can estimate what streams/listeners/saves they need to achieve those PI's.

```
##    Lower Upper StreamsLast28Days ListenersLast28Days SavesLast28Days
## 1:    20    22          2503.057            993.8529        375.5143
## 2:    30    32          9217.862           4097.1724        447.2759
## 3:    40    42         32185.889          13334.5556       1020.4444
##    StreamsLast7Days ListenersLast7Days SavesLast7Days
## 1:          776.500            384.750       67.15000
## 2:         2348.966           1334.759       85.89655
## 3:        10506.778           5801.778      343.00000
```

1.3 Variables plottet vs PI

```
## Warning: Removed 126 rows containing missing values (geom_point).
```



The grafik shows the between 0 and 1 scaled values of the Streams last 28/7 days and listeners last 28/7 days. It can be seen that the Popularity index is dependent in a (not perfect) quadratic function from those variables.

2. Model

This is a regression problem where any model can predict continuous numbers, but the true values can only be of integer type. As an algorithm, the XGboost regression tree was chosen, a (still) state of the art machine learning algorithm that builds on tree boosting.

The model was trained on 263 obversations (75 percent) and fitted on 68 (25 percent) observations (never seen by the model). The test set of 68 data points was sampled at random, but taking into account an equal distribution between new songs (<21 days) and older songs.

2.1 Model performance

```
## [1] 1.524063
```

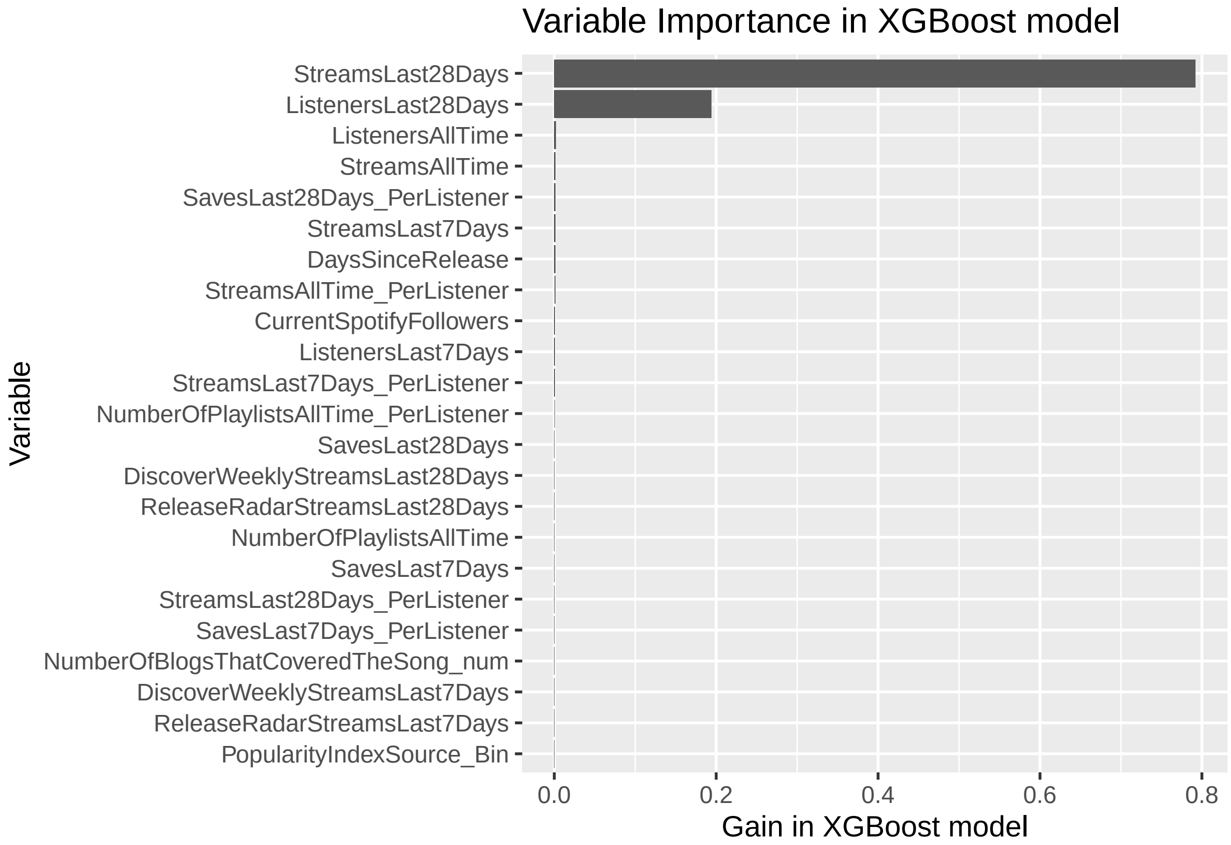
The mean absolute deviation (mad) of 1.52 states that the predictions of the model deviate in mean 1.52 from the true PI. For example if the true PI for one song is 32, we can expect, that the model will predict a value that is around 33.5 or 30.5. So the model is working well but not perfect. There can be several reasons for that the prediction of the model is not perfect.

1. Quality of data: false numbers entered or false PI entered

2. Time-delay of PI (updates every few days, also in Spotify for artist it is not updated in real time)
3. Rounding of PI: Internally the formula from Spotify probably results in a continues number (e.g. 24.32) but is then rounded for the public display. This is sort of imperfect information that can bias the model.
4. There are factors/variables that contribute to the formula that are not publicly accessible/not in the community driven dataset.
5. The algorithm itself lacks potential (model tuning could increase performance or choosing another model like random forest or a neural net)

I personally suspect that it is a combination of 1-3 and think that 4 is rather unlikely.

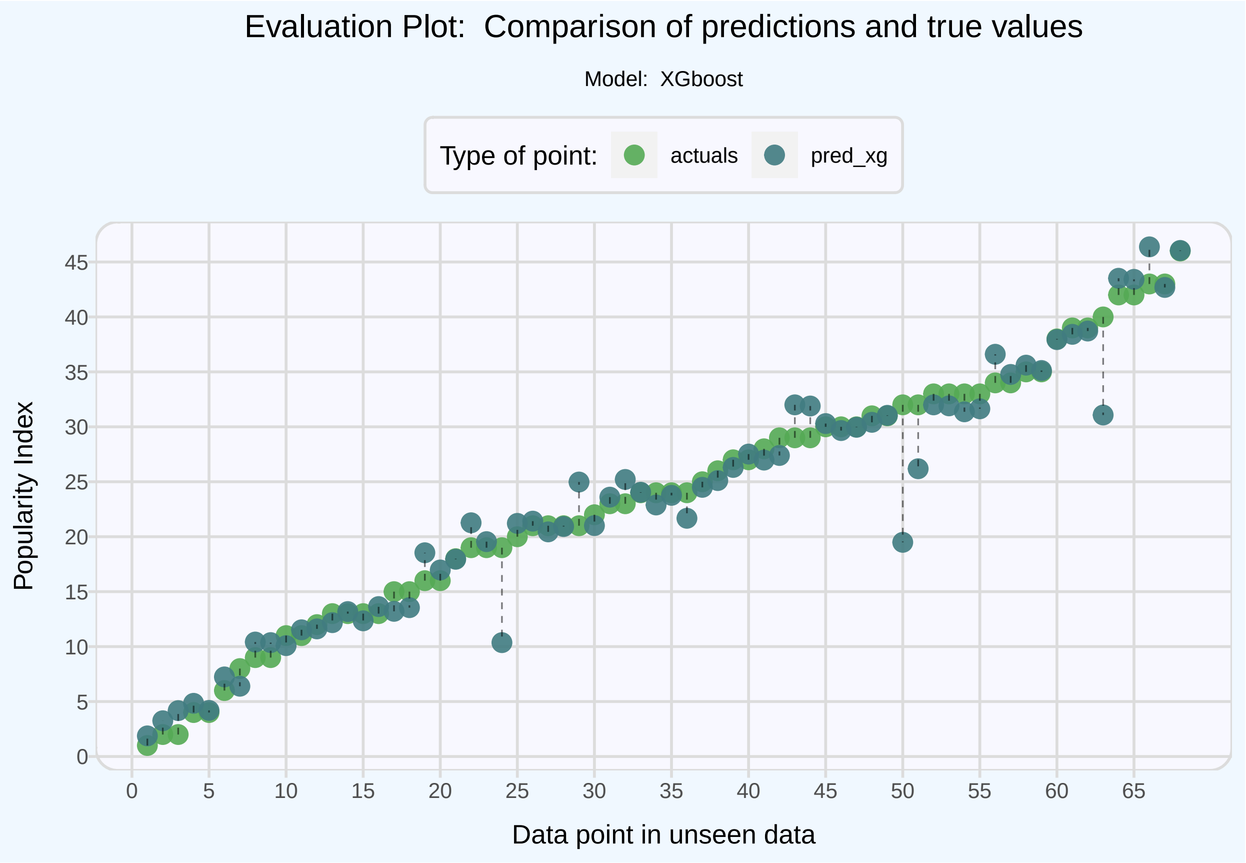
2.3 Variable importance



The grafik shows the variable importance of the computed model. That is, how much each variable contributes to the model. It is seen that streams last 28 days and listeners last 28 days are by far the most important variables. It should be noted that other variables that are highly correlate with StreamsLast28Days and ListenersLast28Days are important too, but the XGBoost model only needs one representor for those sets of variables. It can be observed, that the number of saves do appear very late in the variable importance ranking.

2.3 Graph of predictions

Lets take a look at the predictions and true values of the test data set.



The grafik shows the prediction of the XGBoost model vs the actual values. Most observations were predicted with decent accuracy. Only the predictions of some points show a serious deviation from the actual values. It could be of merit to look into those data points to see which reasons 1-5 from 2.2 do apply.