

Billboard Top 100

Chase Weber

Question

What makes a song successful?

- For over 80 years Billboard has released a weekly Top 100 list to indicate which songs are most popular. What are the features that indicate popular songs on the chart?

Data

- Our data is obtained from data.world
 - <https://data.world/kcmillersean/billboard-hot-100-1958-2017>
- There are two datasets:
 - Billboard Top 100 ratings from 1958 to 2020
 - Song information
- Join these two tables together to achieve a Billboard ranking and general information for 327,895 records.
- 18% of our records do not have song information
 - Impute the missing values with the median value of the songs respective decade.

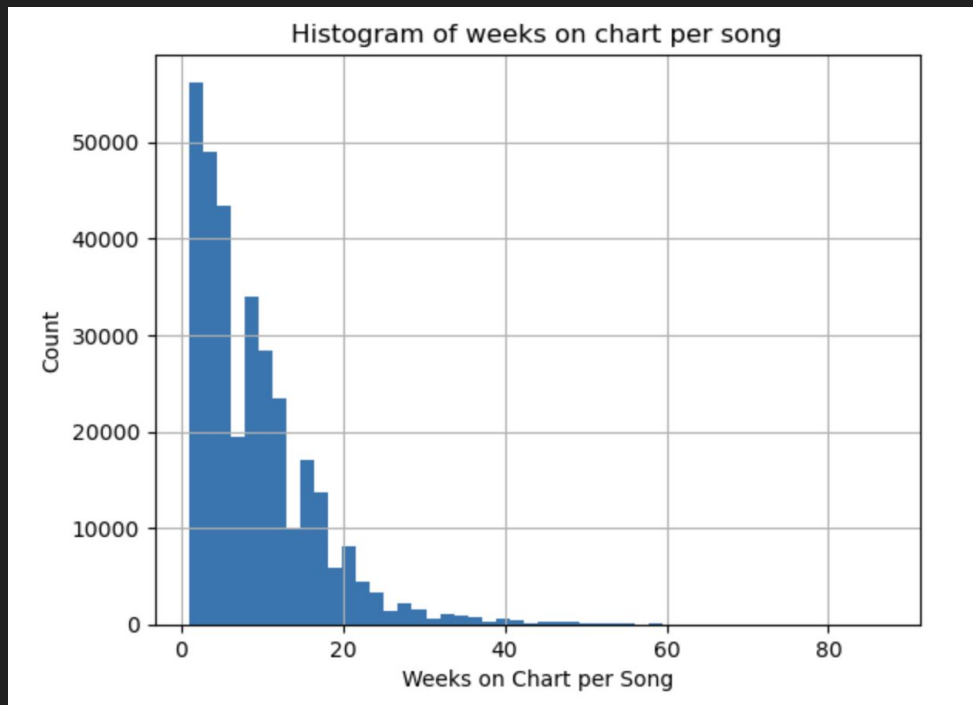
Variable Creation

- **Success Indicator:** if the song was in the top 10 that week.
- **Artist Count Total:** total times an artist appeared on the Billboard 100.
- **Artist Count Year:** times an artist appeared on the Billboard 100 that year.
 - (year prior also calculated)
- **Artist Count Week:** times an artist appeared on the Billboard 100 that week.
 - (week prior also calculated)
- **Emotional Indicators:** did the song score in the 80th percentile of a certain emotion score.

Weeks on Chart Distribution

- Most songs only appear on the chart once.
- Very few make it above 20 weeks.
- The song with the most weeks on the chart: Radioactive by Imagine Dragons, 87 weeks.*

*Note: our data only runs through 2020; as of 2022, the new record is held by Heat Waves, by Glass Animals reaching 91 weeks.



Superlatives (through 2020)

Most sad song: “Sad, Sad Girl” by Barbara Mason

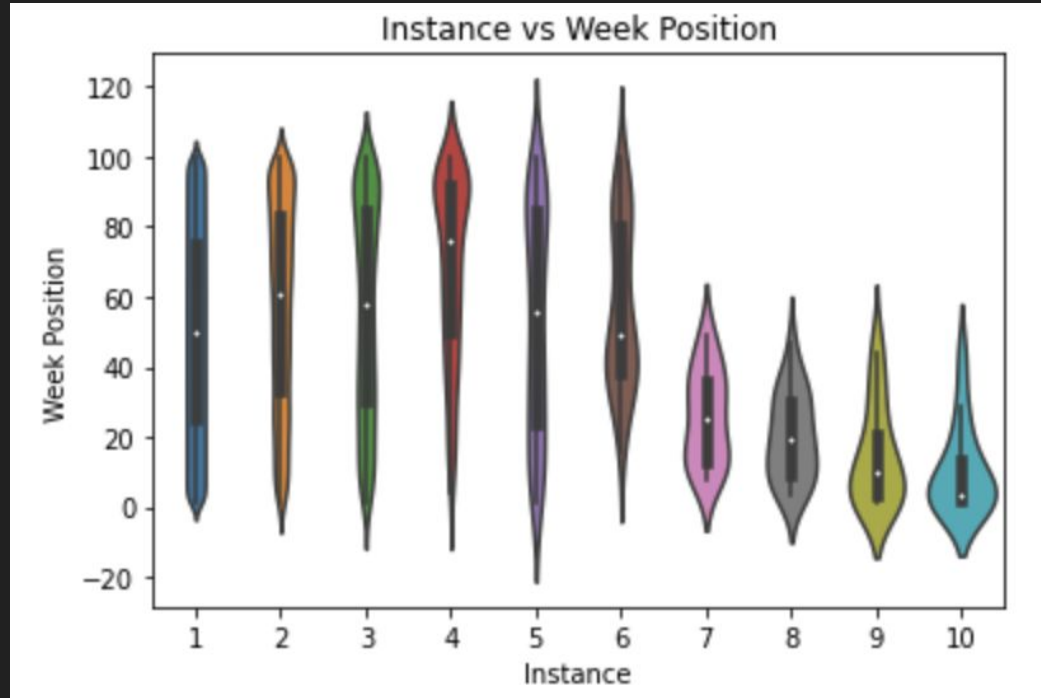
Most joyful song: “I’m in Love” by Aretha Franklin

Most words: “Duckworth” by Kendrick Lamar

Most times on chart (artist): Drake (1,093 counts)

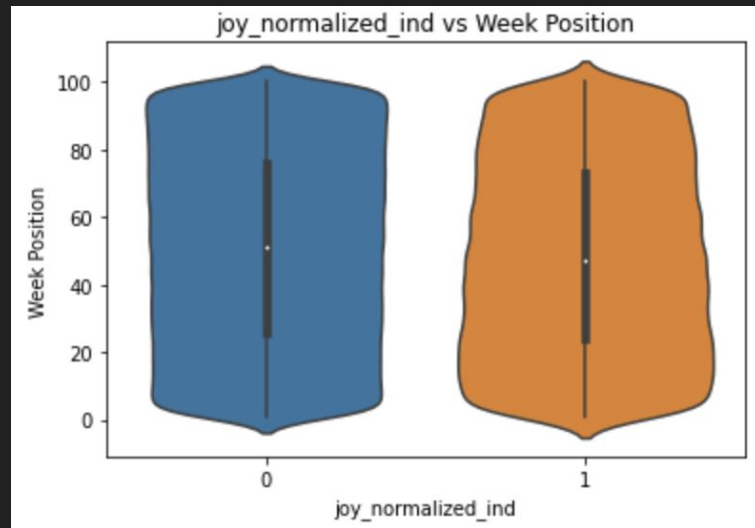
Interesting Relationships

- Instance by Week
Position: instance refers to the amount of times a song has previously entered and exited the Billboard 100. After a song enters its 7th time on the chart, it is almost guaranteed a higher ranking.



Interesting Relationships

- Our “joy” indicator ever slightly appears to be associated with higher ranking songs.
- Weeks on chart as a song stays on the chart longer, the song ranks higher.



Response Selection

- Options
 - Success Indicator: Indicates if a song is in the top 10 that week.
 - Peak Position: The highest ranking a song achieves.
 - Weeks on Chart: The amount of weeks a song remains on the chart.
 - Week Position: The actual ranking of a song.
- Analysis:
 - Using PPS Predictor Analysis, we find that Week Position contains the most predictive factors.
- Response selection : Week Position

Model Comparison

PyCaret iterates model production through our dataset to determine which regression algorithms produce best performing metrics. We find CatBoost, LightGBM, Random Forest, and Extra Trees Regressor to be top performers.

- **PyCaret Analysis:**
 - All Factors

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	5.7055	76.8578	8.7667	0.9079	0.2413	0.2027	7.5700
lightgbm	Light Gradient Boosting Machine	5.7869	78.7547	8.8743	0.9056	0.2406	0.2064	1.3560
rf	Random Forest Regressor	5.9214	82.1829	9.0654	0.9015	0.2434	0.2038	54.5460
et	Extra Trees Regressor	6.0936	86.1485	9.2816	0.8967	0.2490	0.2075	66.0480
gbr	Gradient Boosting Regressor	6.1121	87.5695	9.3577	0.8950	0.2565	0.2267	14.6020

- Controllable Factors:

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	18.1790	523.5849	22.8819	0.3712	0.7147	1.1297	9.3220
rf	Random Forest Regressor	18.2429	523.7808	22.8862	0.3710	0.7144	1.1314	18.1480
dt	Decision Tree Regressor	18.2369	530.7099	23.0370	0.3626	0.7173	1.1306	1.3360
knn	K Neighbors Regressor	19.4559	607.4218	24.6449	0.2705	0.7427	1.1601	641.5360
catboost	CatBoost Regressor	22.3702	689.6900	26.2619	0.1717	0.8279	1.6403	5.0860

Model Comparison

- Having familiarity with LightGBM and Random Forest Regression, these two algorithms are selected for our analysis.
- We use 5-Fold Cross Validation Random Search to tune our model hyper-parameters. LightGBM shows best performance.

LightGBM:

r-square: 0.9015
n_leaves: 1815
n_estimators: 147
max_depth: 3
learning_rate: .82
time_to_run: 1 h 10 m

Random Forest:

r-square: 0.888
n_estimators: 400
max_depth: 7
time_to_run: 10+ h

Model Results

- Recall, we are interested in most influential features (feature importances).
- LightGBM feature importance:
 - Previous Week Position
 - Weeks on Chart
 - Artist Count Year
 - Artist Count Total
 - Artist Count Week
- Random Forest feature importance:
 - Previous Week Position
 - Instance
- All of these features are not necessarily controllable.

Model Reconstruction - Controllable Factors

- Using LightGBM we re-construct our model using only controllable factors.
- Again we use 5-Fold Cross Validation Random Search to tune our model hyper-parameters.

LightGBM:

r-square: 0.369
n_leaves: 666
n_estimators: 88
max_depth: 16
learning_rate: .77
time_to_run: 7 m

Model Comparison - Part 2

- Our top controllable factors are MTLD, Sentiment Score, Word Count, TTR, Emotion Score, and CTTR.
- Notice our metrics, our model with controllable factors does not perform as well as the model with all factors.
- Recall our question - which factors contribute to a song's success?

LightGBM Full Metrics:

MAE: 5.8958
MSE: 79.4554
RMSE: 8.9138
R_Sqr: 0.9044

LightGBM Controllable Metrics:

MAE: 17.9598
MSE: 513.9297
RMSE: 22.67
R_Sqr: 0.382

Recommendations

- A song's position is mostly impacted by uncontrollable factors:
 - The number of times an artist has been and stayed on the chart.
- Looking at controllable factors:
 - Include lyrics with shorter sentences. (MTLD)
 - Create songs with a positive sentiment. (Sentiment Score)
 - Include more words in the song. (Word count)
 - Use repetitive lyrics - less unique words. (TTR)

Future Improvements

- Include additional data in the analysis:
 - External ratings
 - Streaming data
 - Most recent data only (e.g. 2010 and beyond)
- Include additional controllable variables.
 - Song length
 - Marketing information
- Run analysis on all songs, not just those on the chart.

Appendix

- All work can be found on my GitHub:
 - <https://github.com/chase-weber/Billboard-Top-10-Prediction>