DATA 2010

Written Report

Adrien Dinzey, 7842736

Joshua Smallwood, 7826555

Professor: Max Turgeon

Introduction

In the music industry, some artists strive to be on top, having one of the most popular songs out there. One of the most popular ways to track which song stands up on top is the top 100 charts provided by Billboard. Each week, Billboard releases their top 100 songs for that week. As trends change, so do the rankings. Some songs can hold onto a high ranking place for multiple weeks.

The dataset we are analysing for this report consists of two main sources. The first source is a subset of the Billboard top 100 charts, with weeks ranging from 1966 to 2017, and the second is a collection of song features sourced from Spotify. The audio features consist of a range of statistics, in addition to confidence measures on various aspects of the songs.

The analysis question we will discuss in this report is as follows: Can we build a predictive regression model to determine whether a given song will be ranked in the billboard top 100 chart?

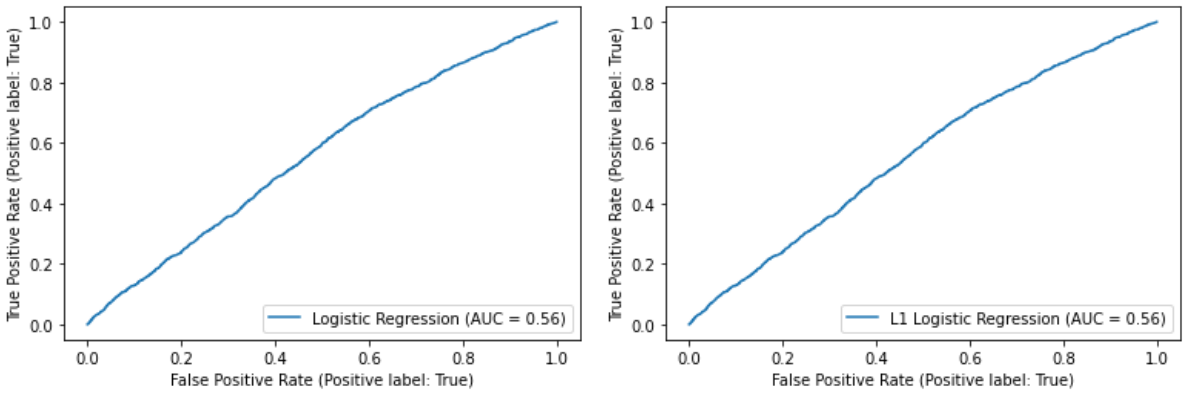
Data & Methods

The first model constructed was a logistic regression model on the predictor variables Danceability, Energy, Valence, Tempo, Track Duration, and Key to determine whether a song will be in the top 10 of weekly rankings. Both standard logistic regression and regularised logistic regression with L1 regularisation were considered. The predictor variables were standardised before training the model, to ensure that the L1 logistic regression model is not affected by the scaling of the data. To measure the performance of the logistic regression models, the Brier score, ROC curves, and precision-recall curves were generated for both logistic regression models.

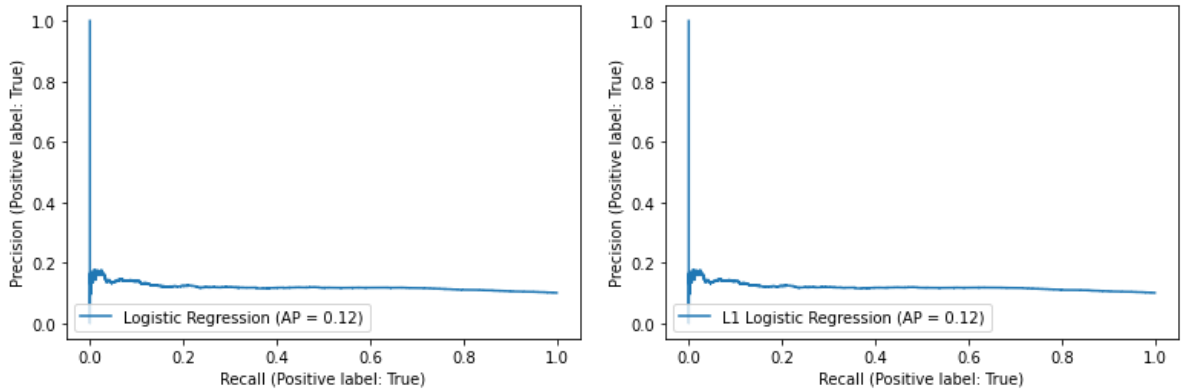
This led us to a follow-up question to our original objective, can we determine the exact spot on the top 100 chart that a song will receive? To accomplish this task, we created a linear regression model. As one would expect, the data varied greatly, mainly due to the amount of variables being taken into account. To combat this, we decided to make a linear regression model and an L2 regularised ridge regression model and compared the goodness of fit of each.

Results

For our logistic regression models, all metrics turned out to be identical, with the models having the same Brier score of 0.101, and having identical ROC and precision-recall curves.



Figures 1 & 2: ROC curves for standard and L1 logistic regression.



Figures 3 & 4: Precision-recall curves for standard and L1 logistic regression

For both models, the AUC is 0.56 and the AP is 0.12.

For our linear regression model, both the standard and regularised models had the same coefficient of determination of $R^2 = 0.004156$ and the coefficients for the variables were almost identical.

```
|Non Regularized Linear Model|
Coefficients:
 [-1.27575263  1.41621646 -0.56418299  0.14405564 -0.0049556   0.14494744]
Root Mean squared error: 28.72

|Regularized Linear Model|
Coefficients:
 [-1.27574734  1.41620885 -0.56418017  0.14405694 -0.00495486  0.14494693]
Root Mean squared error: 28.72
```

Figure 5: Linear regression coefficients for both models

As such, we continued with our analysis by choosing the ridge regression model. We see that the RMSE is 28.72. Considering the possible values of [0,100], this is a very large RMSE, indicating that the model is not very effective. We then examined each variable's effect on the predicted outcome (with all other factors considered) to explore what we could. In the below figures, the observed weekly positions are plotted in black, while the model's predicted positions are plotted in blue.

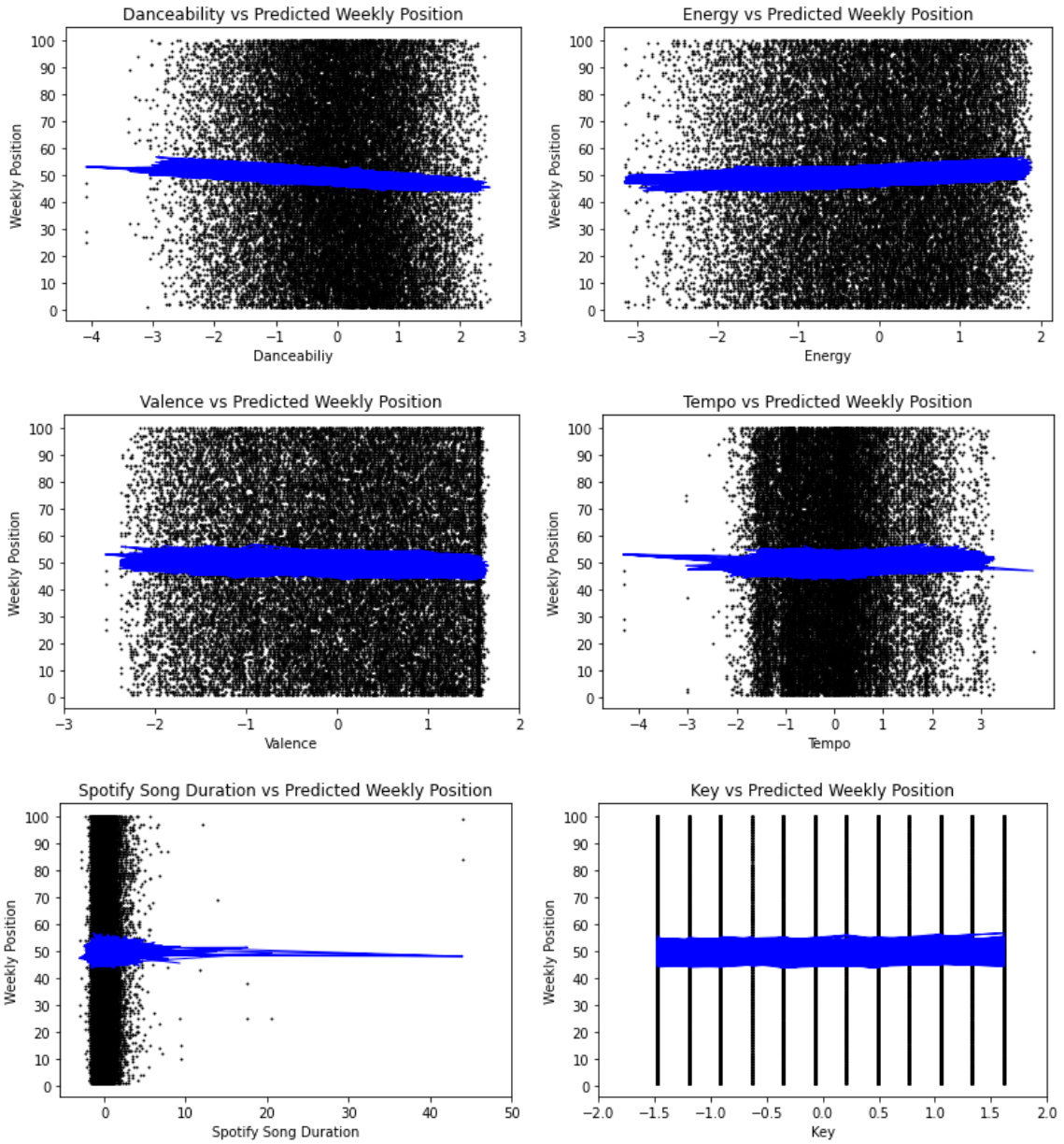


Figure 6: Predicted weekly position vs. input values for each input variable

It appears that the actual predicted values vary greatly from each other, and so the linear model's predicted values, while trying to find an average weekly position, if tested against real, varied, data it will yield high error; as observed.

Discussion

For the logistic regression models, the Brier score is low, at around 0.1, implying that the predictions are well calibrated. However, the ROC and precision-recall curves indicate that the model is not appropriate. The ROC curves are not close to the upper left corner, and the AUC is close to 0.5, meaning that the classification is not very good. The precision-recall curve is not close to the horizontal line precision = 1, meaning that the model does not have a high level of recall.

For the linear regression models, the very low coefficient of determination and high RMSE indicates that the linear model is very unreliable and fails to accurately model the data. This indicates that there are likely more variables that determine the success of a given billboard top 100 song. This indicates that these variables do not work linearly together, meaning by modelling them as strictly increasing and decreasing in tandem we will not reach a useful model. As a result of this, it is possible that there are perhaps “sweet spots” for these variables instead. That is, there are certain values or ranges for each variable where a song will be most likely to be higher on the top 100 chart. For further analysis, we could try different types of regressions such as polynomial regression or using a generalised linear model, with distributions in the exponential family.

There are many factors that are extremely difficult to analyse, for example the time era. It is well known that popular culture has gone through many phases through the years/decades and with it so has popular music. Another variable that could be explored in future analysis would be the success rate of these variables when grouped into different time periods. So for example, 80’s music might appreciate energetic, danceable music. While the 70’s might have a greater appreciation for happier, relaxing music. While these factors might lead to more success, there is still the inevitable hurdle of differing music tastes. No matter which time period, there will always be lots of variation as the people of the world have varying music tastes.