Multivariate Data Analysis Project

OK Computer

Is it possible to predict a song's genre
and popularity based on its lyrics?

Lili Sára Sarkadi-Nagy
Corvinus University of Budapest
Year 2

Contents

1 Introduction

We all have an intuition about the fact that certain musical genres are characterized by a particular vocabulary. As they have different origins and cultural influences, I hypothesized that it is possible to predict a song's genre and popularity based on its lyrics. Are there certain words in some lyrics that contribute to their success? Is the vocabulary of certain lyrics quantifiable and based on that, is it possible to predict the genre? In the recent decades, machine learning and text mining have been gaining momentum and are used together in many cases, like analyzing and categorizing reviews, comments, forums etc. Combining these two methodologies by quantifying qualitative data can answer these questions. The codes of the analysis are available on Github.

2 Data

For the analysis, I used the Genius expertise dataset from the analysis of Lim and Benson (2020). They scraped Genius, a crowd-sourced lyrics and musical knowledge website for their network-focused analysis, and published multiple datasets online. In this analysis, two databases were merged based on the song identifyers - *song_info* for the metadata and *lyrics* containing the uploaded lyrics to the website.

2.1 Data cleaning

To prepare the data for quantitative analysis, I preprocessed the lyrics by removing punctuations, stop words, lower casing, stemming and lemmatization. The dataset contained non-English lyrics, which were removed. Afterwards, I specified a vectorizer and created a sparse matrix from the words that appear in more than 10% but in maximum 80% of the lyrics, and filtered out additional filler words manually. Overall, the result was a bag of words matrix composed of 410 variables and over 34000 observations. Afterwards, I simplified the genres of the songs by only including the so-called main tags on the website (Pop, Rock, Country, R&B, Rap). Since the uploaded content is Hip-Hop heavy and also due to computational capacity constraints, I randomly selected approximately 3000 songs the genre-composition of which are better suitable for analysis. The number of pageviews a song gets is also attached to the dataframe.

2.2 Methodology

There are numerous reasons why statistical methods that do not involve any machine learning are not appropriate for such data. The statistical science occupies itself with the estimation causal and non-causal inference based on hypotheses tested using probability models (Ij, 2018). Machine learning is also applied to a set of specific problems problems, which sometimes overlap with statistics. A key difference between the two fields is for instance that statistical methodologies best work on a limited number of variables. On the contrary, machine learning methods can best solve problems where the focus is on pattern recognition and prediction with a large number of variables even in cases when they may be difficult to interpret individually. Textual analyses have been gaining momentum in the recent time and since the nature of text data is usually unstructured, machine learning algorithms are canonically used to classify and interpret such data (Khan et al., 2010). This is why I decided to compare and evaluate machine learning techniques for the categorization of genres of songs and the prediction of views on the songs.

3 Categorization

3.1 Decision tree

To determine which category a song belongs to, the most suitable tool is a classification tree. There are more than 400 explanatory variables which are sparsely distributed which can be handled by this algorithm, however, the results can end up to be biased. Based on my sample, the regression tree discarded Country and Rock songs due to their low overall proportion, and the most important nodes are represented on Figure 1.

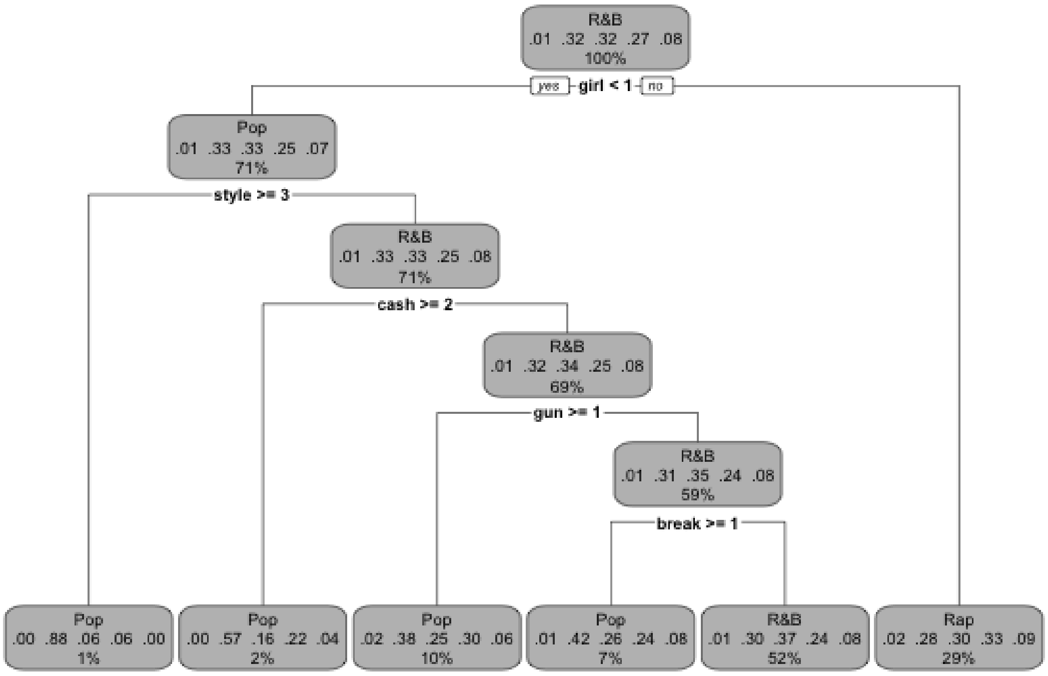

Figure 1. Decision tree

The overall depth of the first estimated tree is 47, the graph only represents the most important ones, and the tree has to be pruned to avoid overspecification. The accuracy of the tree based on the confusion matrix is 0.31, which is better than randomly selecting a category out of the five. Some of the most important words in predicting a song's genre were *girl, style, cash, gun, break, shawty, money and block.*

3.2 Boosting algorithm

Boosting algorithms can be used to improve on prediction quality by producing a series of k weak models and improving them iteratively. The AdaBoost is an algorithm especially designed for classification problems, like this one. It weighs mis-classified data points and each such step is reflected in the next iteration. In this case, it produced a completely different result from the previously described tree.

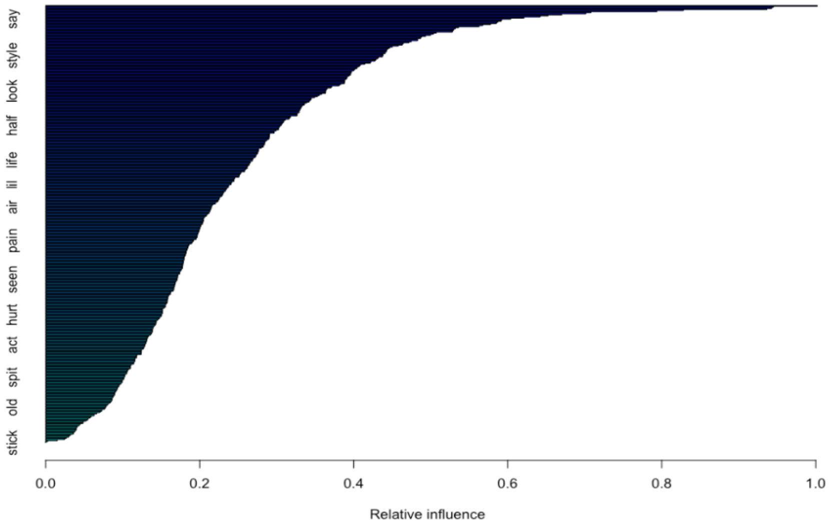

Figure 2. Relative importance of words in genre classification - boosting algorithm

The most important variables here in predicting a song's genre turned out to be (in this order): *stay, love, know, real, girl, beat* and *feel.*

4 Popularity prediction

My additional hypothesis is that the lyrics have an explanatory power in the popularity of a song. As the songwriters try to be relatable to the audience, the lyrics of a hit have to be simple, comprehensible and commercial, I expect that the lyrics of popular songs have some characteristics in their vocabulary in common, especially if the data is segmented by genre. This is also helpful because of computational capacity constraints.

4.1 Stepwise model selection

First, I estimated a OLS regression where the explanatory variables are the sparse matrix of the words, and the dependent variable is the number of pageviews a song received. The *no of views* variable is highly skewed, therefore I will take its logarithm. The R^2 of the regression is quite low (0.02). The next step in a model specification problem would be backward elimination - omitting the insignificant variables from the regression. However, with this amount of variables this is not possible: therefore I used the stepwise model selection algorithm, which is based on the Akaike Criterion. The AIC can be estimated as

$$AIC = 2k - 2ln(L)$$

where k is the number of parameters and L is the maximized value of the likelihood function. The algorithm compares the possible model specifications and picks the most optimal one based on the tradeoff between goodness of fit and information loss. One drawback of this method is its complexity, as the input size (in this case, the number of variables) increases, the time it takes the algorithm to produce a result increases exponentially, as the number of operations as a function of input size can be written as a geometric series ($2^n - 1$), the number of equations the AIC it needs to evaluate with 410 explanatory variables would be $2^{410} - 1$.

4.2 Random forest

Machine learning algorithms can be used to estimate a model to test this hypothesis. Random forest models construct multiple decision trees, and by randomizing predictors and using bootstrap aggregation, it achieves a high level of accuracy. (Breiman, 2001) Using this algorithm, I got the result that the variance in the lyrics of the song account for 1%, 3%, 0.7% and 3% of variance in the popularity of Rap, Pop, RnB and Rock songs. There were too few non-empty observations to apply the same method to Country songs. These are similar values to the ones in the estimates of the previously mentioned OLS regressions.

5 Conclusion

My initial hypotheses were that on a song's lyrics, it is possible to predict its genre and popularity. To answer these questions, I used data from genius.com, and due to the large number of variables and observations, I decided to use and compare machine learning methods such as decision trees, boosting and random forest. With the AdaBoost algorithm, it is possible to predict a song's main genre (out of five possibilities) with 30% accuracy. However, the results for the popularity estimate suggested that based on the number of certain appearing words it is hardly possible to foresee how popular a song will be, further scope of this investigation could be sentiment analysis where the lyrics are inspected as a whole unit, rather than the sum of words. Further limitations were the number of included words, which could be further extended, and the exclusion of subgenres, which would provide a more detailed picture.

Bibliography

Breiman, L. (2001). *Random forests.* Machine learning, 45(1), 5-32.

Ij, H. (2018). *Statistics versus machine learning.* Nat Methods, 15(4), 233.

Hothorn T, Zeileis A (2015). *partykit: A Modular Toolkit for Recursive Partytioning in R.* Journal of Machine Learning Research, 16, 3905-3909. https://jmlr.org/papers/v16/hothorn15a.html.

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). *A review of machine learning algorithms for text-documents classification.* Journal of advances in information technology, 1(1), 4-20.

Kuhn M, Wickham H (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* https://www.tidymodels.org.

Lim, D. & Benson, A. (2020) *Expertise and Dynamics within Crowdsourced Musical Knowledge Curation: A Case Study of the Genius Platform.*

Milborrow, S. (2016) *rpart.plot: Plot rpart Models. An Enhanced Version of plot.rpart..* R Package.