MACHINE LEARNING 1
JUNE 2020

EXPLICIT CONTENT PREDCITOR IN SONGS

PROJECT REPORT

Elías Abad Rocamora
Marc Fuentes i Oncins
Alex Martí Guiu
Barcelona, UPC - FIB, FME & ETSEIB

Index

Introduction

It is curious how the language has some expressions or words marked as “not suitable for every audience”, this expressions help emphasize the language but their usage is usually rude or shocking. Lately, explicitness in music is rising and more and more explicit songs are getting o the top. But, do these expressions affect the musical properties of a song itself?

Our objective is to predict if a song uses explicit content based on its instrumental features. We, as music consumers (we all consume music daily), are motivated to find out if there are correlations between the variables that we can extract (song features), and in specific, the explicitness of the songs. We want to know if “popular songs” with explicit content in it can be statistically proven as different from clean songs based on their musical features, as well as the relations and correlations between this features.

To do this we will use a dataset of the top 100 Billboard songs from each week since 1958. It shows 22 attributes related to the songs. From the author, song name or genre to other technical aspects like tempo, loudness or danceability.

After having developed our models, we will want to check our results by applying the predictions in songs that we are interested in (taking into account that the study will be performed on popular songs, therefore the result should, in theory, be more accurate if the song is well known). We will use the spotify “Audio Features for Tracks” tool for developers, with which we will make tests for songs we like or that we are looking forward to analyze.

1. The dataset

Our data set contains 22 parameters but for the analysis we are interested in the last 15 ones corresponding to:

1. Explicitness - *factor*
2. Duration (in ms)
3. Popularity (between 0 and 100)
4. Danceability (between 0 and 1): an estimation based on other parameters about how danceable a song seems to be.
5. Energy (between 0 and 1): a value describing how energetic a song is.
6. Key: in which key is the song made (0 corresponds to C, and so on for the following keys in musical scale).
7. Loudness (in dB): the average loudness of the song, usually varies from -60 to 0 dB.
8. Mode - *factor* : Mode indicates the modality (major or minor). Major is represented by 1 and minor is 0.
9. Speechiness (from 0 to 1): a measure of spoken words or spoken-style singing found in the song (rap → 1, melodic → 0).
10. Acousticness (from 0 to 1): how acoustic (not electronic) a song is.
11. Instrumentalness (from 0 to 1): a measure of the amount of vocals a song has.
12. Liveness (from 0 to 1): probability that the song was recorded live.
13. Valence (from 0 to 1): measures the positivity of a sound, if it is happy, cheerful, euphoric etc.
14. Tempo (in Beats Per Minute)
15. Time signature: beats per bar of the song.

2. Missing data

The dataset that we've chosen isn't complete (like most datasets), there are many rows with missing parameters. Most of the rows with missing parameters had many of them, almost every time a column is missing, the rest of them are also missing. This is why we have chosen to delete the rows where the target variable (explicit) or "danceability" are missing.

3. Exploration process

In a first approach to understand the data we are using we will extract some information before the data treatment.

We are going to set some *a priori* hypothesis from simple histogram based analysis from the pre processed data to see the difference between the distribution of these variables on explicit vs non explicit.

By seeing the distribution of the initial data we can estimate some possible outcomes and detect singular distributions in the data. In the following plots the red bars correspond to explicit songs and the green to non-explicit.

When observing the danceability histogram (*see fig.1*) it can be seen that the explicit songs tend to a higher level of danceability centered around 0.75 while non-explicit songs is centered at 0.6.

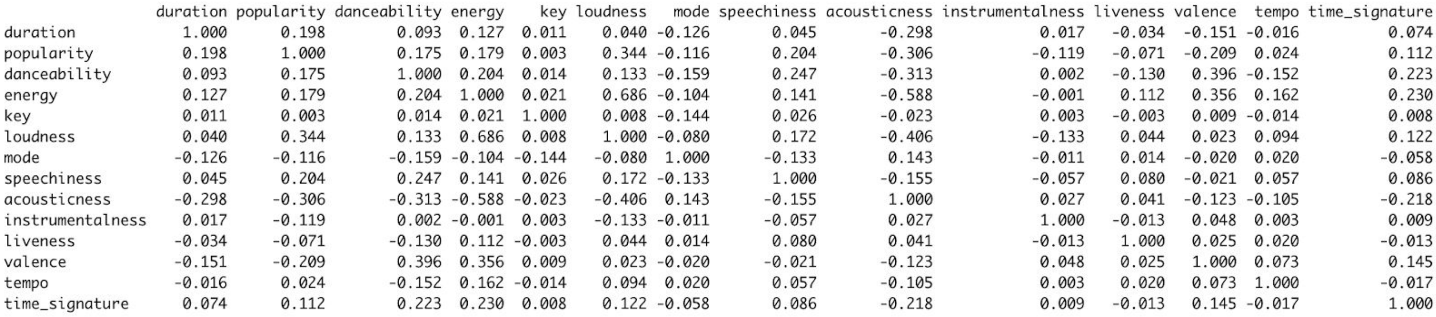
Looking at the popularity histogram (*see fig.2*) it can be seen that explicit songs tend to be more popular than non explicit ones. In the non-explicit case the popularity seems uniform at the beginning but then tends to zero after passing the value 60. In the other case the explicit songs shows kind of a gaussian curve centered at 65 with a high concentration of values between 50 and 80.

Looking at the data from the key values (*see fig.3*) we see that there is a general trend, specially in the explicit songs for the 0 key (C). The remaining data shows a more or less uniform distribution with a minimum on the 2nd key.

Looking at the loudness of the data (*see fig.4*) it can be seen that non-explicit songs are less loud in general but have a greater variance. Explicit songs are louder and use a smaller range of values.

Regarding to the speechiness of the songs (*see fig.5*) we see that in the non-explicit case, songs are mostly non-spoken at all. The explicit songs are widely distributed over the axis which can be thought as urban songs that usually use a stronger language and have a greater speechiness index.

We will also take a look at the correlation matrix of the data.



```
                 duration popularity danceability energy    key loudness   mode speechiness acousticness instrumentalness liveness valence  tempo time_signature
duration            1.000      0.198        0.093  0.127  0.011    0.040 -0.126       0.045       -0.298            0.017   -0.034  -0.151 -0.016          0.074
popularity          0.198      1.000        0.175  0.179  0.003    0.344 -0.116       0.204       -0.306           -0.119   -0.071  -0.209  0.024          0.112
danceability        0.093      0.175        1.000  0.204  0.014    0.133 -0.159       0.247       -0.313            0.002   -0.130   0.396 -0.152          0.223
energy              0.127      0.179        0.204  1.000  0.021    0.686 -0.104       0.141       -0.588           -0.001    0.112   0.356  0.162          0.230
key                 0.011      0.003        0.014  0.021  1.000    0.008 -0.144       0.026       -0.023            0.003   -0.003   0.009 -0.014          0.008
loudness            0.040      0.344        0.133  0.686  0.008    1.000 -0.080       0.172       -0.406           -0.133    0.044   0.023  0.094          0.122
mode               -0.126     -0.116       -0.159 -0.104 -0.144   -0.080  1.000      -0.133        0.143           -0.011    0.014  -0.020  0.020         -0.058
speechiness         0.045      0.204        0.247  0.141  0.026    0.172 -0.133       1.000       -0.155           -0.057    0.080  -0.021  0.057          0.086
acousticness       -0.298     -0.306       -0.313 -0.588 -0.023   -0.406  0.143      -0.155        1.000            0.027    0.041  -0.123 -0.105         -0.218
instrumentalness    0.017     -0.119        0.002 -0.001  0.003   -0.133 -0.011      -0.057        0.027            1.000   -0.013   0.048  0.003          0.009
liveness           -0.034     -0.071       -0.130  0.112 -0.003    0.044  0.014       0.080        0.041           -0.013    1.000   0.025  0.020         -0.013
valence            -0.151     -0.209        0.396  0.356  0.009    0.023 -0.020      -0.021       -0.123            0.048    0.025   1.000  0.073          0.145
tempo              -0.016      0.024       -0.152  0.162 -0.014    0.094  0.020       0.057       -0.105            0.003    0.020   0.073  1.000         -0.017
time_signature      0.074      0.112        0.223  0.230  0.008    0.122 -0.058       0.086       -0.218            0.009   -0.013   0.145 -0.017          1.000
```

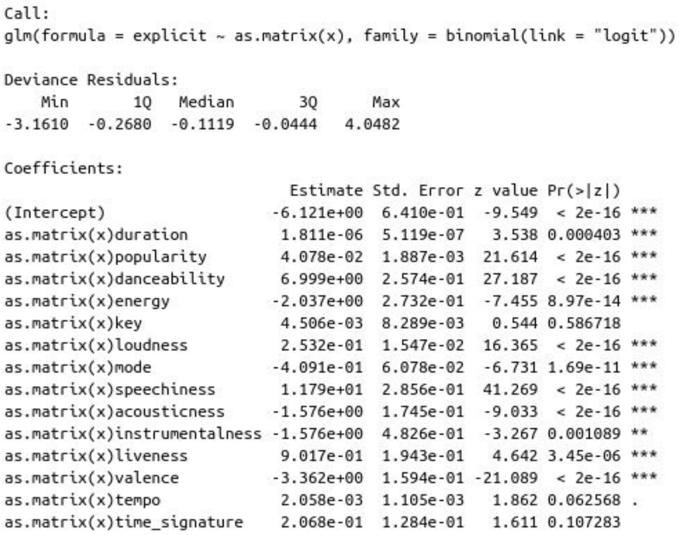
If we take a closer look we can see that the variables are in general low correlated but there are some special cases where the variables seem to have some kind of correlation. We can see this between the variables energy and loudness with a correlation value of 0.686 and energy and acousticness with -0.588. It is logical to think that, as shown, energetic songs are loud and tend to be more electronic and less acoustic.

4. Variable selection

In order to reduce the amount of variables we are using we will define a first GLM of the data. By doing this we can get a first raw approach of how the variables behave corresponding to their effect on the explicitness of the songs.

For our following models and tests we will be using an alpha value of 90%.

We have defined the following model:



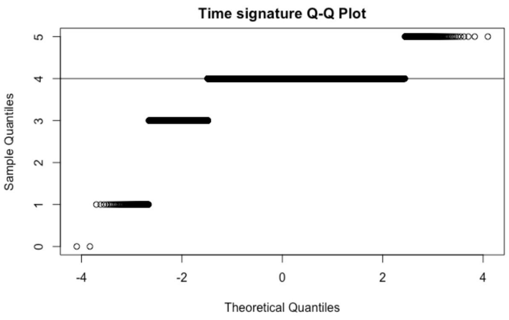
```
Call:
glm(formula = explicit ~ as.matrix(x), family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1610  -0.2680  -0.1119  -0.0444   4.0482

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -6.121e+00  6.410e-01  -9.549  < 2e-16 ***
as.matrix(x)duration          1.811e-06  5.119e-07   3.538 0.000403 ***
as.matrix(x)popularity        4.078e-02  1.887e-03  21.614  < 2e-16 ***
as.matrix(x)danceability      6.999e+00  2.574e-01  27.187  < 2e-16 ***
as.matrix(x)energy           -2.037e+00  2.732e-01  -7.455 8.97e-14 ***
as.matrix(x)key               4.506e-03  8.289e-03   0.544 0.586718
as.matrix(x)loudness          2.532e-01  1.547e-02  16.365  < 2e-16 ***
as.matrix(x)mode             -4.091e-01  6.078e-02  -6.731 1.69e-11 ***
as.matrix(x)speechiness       1.179e+01  2.856e-01  41.269  < 2e-16 ***
as.matrix(x)acousticness     -1.576e+00  1.745e-01  -9.033  < 2e-16 ***
as.matrix(x)instrumentalness -1.576e+00  4.826e-01  -3.267 0.001089 **
as.matrix(x)liveness          9.017e-01  1.943e-01   4.642 3.45e-06 ***
as.matrix(x)valence          -3.362e+00  1.594e-01 -21.089  < 2e-16 ***
as.matrix(x)tempo             2.058e-03  1.105e-03   1.862 0.062568 .
as.matrix(x)time_signature    2.068e-01  1.284e-01   1.611 0.107283
```

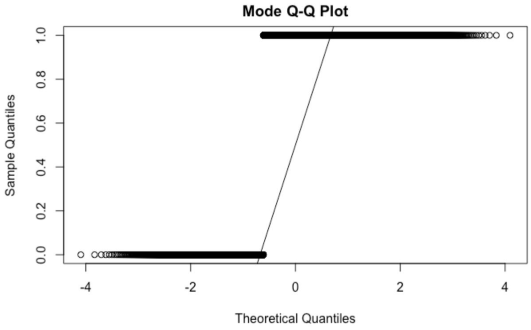
Looking at the summary of the model and focusing on its ANOVA test we can see that the variables key and time signature are not significant to the model. It is logical to think that the key in which the song is made does not imply the existence of explicitness and that the time structure of the song doesn’t affect the resulting language used. We can see that variables like danceability and popularity, among others, are strongly significant.

This is not our final model, yet it has given us important information.

Another important thing to see is the data distribution. We assume that the data follows a normal distribution but we need to check this for all variables.

With the *qqnorm* and *qqline* functions in R we can see how well the data adjusts compared to a normal distribution. We have tested this on all variables and we have seen two special cases.





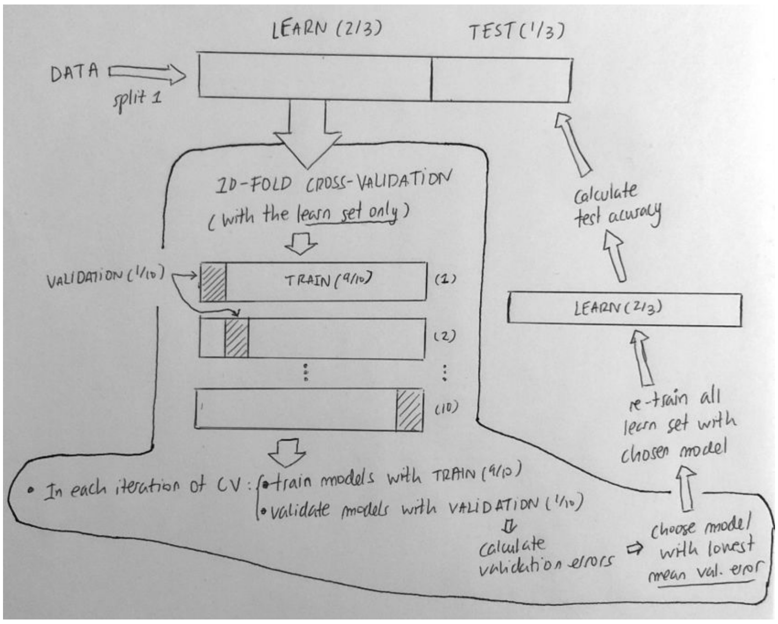
The variables time signature and mode differ from the normal distribution. In the mode case, our variable only takes 0 or 1 values with a preference on 1. It is clearly not normal but it is a factor that can be applied later on and can be useful, furthermore the ANOVA test proved its significance.

For the other variable, time signature, we can see how the values do not correctly adjust to the normal distribution. In addition to the fact that it was not significant this variable is not suitable for our project.

From now on we will not work with the variables key and time signature.

5. Training the dataset

We are going to consider a k-fold cross-validation to choose the model that best fits our data to then use it to classify the output to decide whether a song is explicit or not. To do so, we will first split the data in two sets, the learning one which has ⅔ of the songs and will be used to apply the 10-fold cross-validation to choose the best model, and the test set which has ⅓ of the songs and will be used to test the goodness of fit (test accuracy) of our final classifications with the final model chosen by the cross-validation which will have previously been trained with all the ⅔ of learn data.


LEARN (2/3)
TEST (1/3)
DATA
split 1
10-FOLD CROSS-VALIDATION
(with the learn set only)
VALIDATION (1/10)
TRAIN (9/10)
(1)
(2)
(10)
Calculate test accuracy
LEARN (2/3)
re-train all learn set with chosen model
• In each iteration of CV: • train models with TRAIN (9/10)
• validate models with VALIDATION (1/10)
Calculate validation errors
Choose model with lowest mean val. error

These are the models that we will use in the cross-validation, from which we will choose the best one to train the whole learn set with it:

a) Model 1 → Classification using a two-group Dichotomizer, with the only assumption that the data is normal, which we have proven to be for most explanatory variables as we have seen with the *qqplots*. The dichotomizer is built with the following discriminant functions for each group k:

$$g_k(x) = ln\{P(w_k)p(x|w_k)\} = ln\,P(w_k) - ln\{(2\pi)^{d/2}|\Sigma_k|^{1/2}\} - 1/2(x-\mu_k)'\,\Sigma_k^{-1}(x-\mu_k)$$

and defining the dichotomizer as $g(x) = g_1(x) - g_2(x)$, where a positive result means the result of the classification is explicit and otherwise for the negative case.

This first model would be the equivalent to applying QDA classification (R function) but we have decided to implement it ourselves to understand better the process behind.

We learned a lot by doing this because we spent a whole afternoon looking for errors in the code and it was only a parenthesis.

b) Model 2 → Regression through a GLM using the binomial family (1 → explicit, 0 → non explicit) with link *logit* with witch we will use it's predictions to classify to both groups using the *round* function (we get numbers between zero and one, and classify to the closest group to the prediction).

c) Model 3 → We will consider using a third model of classification using k-nearest-neighbours (knn). In order to speedup our program we will fix our k-value before executing the model selection (10-fold cross validation). Using the full learning data and contrasting with the test set we have found the best value (least error) to be k=11.

d) Model 4 → We will build a Neural Network, specifically a Multilayer Perceptron. First we have found a value for the decay of the regularization, fixing the size to 20, by using 10x10CV with all the learn set. We have obtained that the best decay for our neural network is 0,4. Therefore our NN will end up being 12-20-1 regularized with decay 0,4. It has 12 entries as we have 12 explanatory variables, which ends up resulting in 281 in total for the MLP.

6. Results obtained

After the 10-fold cross validation, the first three models gave almost the same results in terms of mean validation errors (~ 7%) and F1 scores (~96%) and the last one seems to have a higher error.

Model	CV Mean Error	CV Mean F1 score
Model 1 (qda dichotomizer)	0.0705	0.9606
Model 2 (glm binomial)	0.0690	0.9619
Model 3 (11-nn)	0.0677	0.9625
Model 4 (MLP)	0.1031	NA

Note: The F1 score isn't available for the output of the MLP as it sometimes classifies all observations as non-explicit.

The variances of the errors

Model	Error Variance	F1 Variance
Model 1 (qda dichotomizer)	5.519e-05	1.722e-05
Model 2 (glm binomial)	7.214e-05	2.253e-05
Model 3 (11-nn)	3.571e-05	1.159e-05
Model 4 (MLP)	1.593e-04	NA

We asume that the mean and the variance are asymptotically normal (if we had a big enough sample we could apply the Central Limit Theorem), so we are going to test if the mean errors of each model are significantly different.

To see this, we are going to test if the difference in the mean of errors in every iteration of the c.v. is equal to 0. We can do that with the R function t.test.

These are the results:

Hypothesis	p-value
Error_dichotomizer = Error_glm ?	0.6767
Error_dichotomizer = Error_11-NN ?	0.3703
Error_dichotomizer = Error_MLP ?	5e-06
Error_11-NN = Error_glm ?	0.7061
Error_11-NN = Error_MLP ?	3e-06
Error_glm = Error_MLP ?	2e-06

So we can conclude that there’s no significant difference between the validation error of the first three models as al p-values are > alpha = 0,05. Also, the last model (MPL) is significantly worse than the other ones as the p-values are always < alpha = 0,05.

7. The Decision - Model Choice

In terms of error rates and variances we have seen how the best model is *knn* but by a small difference. Because the three models are so similar, if we wanted to minimize the complexity of the model, we would choose the second model (logistic regression).

Another fact that leads us to choose the GLM as the best one, is that this kind of models are easily interpreted and give us a mathematical definition of the data, while in KNN, the training data is the model in the sense that we compare a new observation with the ones we know to make a choice in the class.

Models 1 and 2 are similar in the sense that they are both interpretable models and have almost the same accuracy, but we consider that logistic regression is better because it is less complex (less parameters).

Therefore, we will proceed to train our whole learning dataset (⅔ of the total) with the chosen model, the GLM Regression, and test its accuracy with the test set. Furthermore, once we have the model trained, we will be able to download track features from the Spotify Developer API (url in *References*) from which we can acquire the explanatory variables of our data and therefore use the already trained and tested model to classify that song.

8. Test with specific songs

Now that we have objectively chosen a model, we can proceed to check the prediction results and try to explain some errors, as well as make predictions for songs which weren't in our dataset. To explain the errors, we have selected a few songs that were wrongly predicted by our model and try to find an explanation to understand some possible future behaviours.

8.1 Analysis of Errors

Here are some songs that are predicted wrongly by all the models:

- *China*, by Anuel AA - (obs 155)
- *Let it be,* by The Beatles - (obs 3625)
- *Hey Ya!*, by OutKast - (obs 8069)
- *Enemies*, by Post Malone - (obs 5)
- *EARFQUAKE*, by Tyler the Creator - (obs 100)

The first three songs are predicted as explicit when they are really not:

- *China* is an urban latin song, which is a genre that presents explicitness in their songs in a regular basis. The musical properties match the reggaeton/urban style and its high popularity makes the model predict it as explicit.

- *Let it be* is a very famous Pop-rock song by The Beatles. In this genre explicitness isn't commonly found and this prediction is really strange. If we take a closer look at the song we can see that it was very popular, with a value of 80. It also has a loudness of -8 dB which in the variable histogram can be seen that corresponds to more explicit songs than non-explicit. This could be the reason to the wrong prediction by the model.

- *Hey Ya!* is a well known hip-hop/rap song. It presents high values of popularity, danceability and energy which are more likely to be found in explicit songs. It also presents a really high value of valence, meaning it's a positive and happy song which can be easily seen by listening to it. The high popularity, danceability and energy makes the model tend to the explicit prediction.

The other two songs are predicted as non-explicit when they are explicit:

- *Enemies* is a hip-hop/melodic rap song. When taking a closer look at the musical features we can see it has a danceability of 0.54, a low value for an explicit song. The other values if compared to the histograms are in zones of equal distribution for both classes. The low danceability might be the determining factor for the classifier that ends up choosing it not to be explicit.

- *EARFQUAKE* is a hip-hop/rap song that presents different musical properties when compared to most pieces of the genre. It has a low danceability and energy, and a high value of acousticness. This features tend to be in non-explicit songs and as the model doesn't take the genre into account it can't recognize this song as "rap" which is usually associated with strong language.

8.2 Test of our predictor with new songs

Once seen some kind of special cases that our model fails to predict we can try some other songs which were not present in our dataset, that have unconventional combinations of features and genres or that are popular at the moment. We have extracted all their features and variables that our model uses from the Spotify for Developers API, which returns us a .json file with all the features. The only feature that we need and Spotify doesn't give us is "Popularity", which we have injected manually according to the song's popularity at the moment (based on the number of Spotify reproductions of the song)

We have chosen the following songs:

- *Pigs (Three Different Ones)*, by Pink Floyd
- *Say so*, by Doja Cat
- *ROCKSTAR*, by DaBaby
- *Sunflower*, by Post Malone

The first song is identified as explicit by spotify which seems strange considering it is from a Classic Rock album of 1977. When testing it with our model we predict that it is non explicit, which is wrong but we consider it could be a human mistake too as we think that it doesn't have much explicit content in it and the classification made by spotify is quite surprising.

- Prediction for Pigs, Pink Floyd: 0.01943786 , rounding, we get: 0 (non explicit)

When looking at the second song, Say So by Doja Cat, we see an R&B piece where the instrumental emits optimism and good vibes that has been very popular in the last year. The song is flagged as explicit. When testing it with our model, we get an accurate prediction of explicitness of the song.

- Prediction for Say So, Doja Cat: 0.7751583 , rounding, we get: 1 (explicit)

For the third song, *ROCKSTAR*, we have a Hip-Hop/Rap song which is the number 1 song on the Spotify global list (as per june 2020). This genre is more likely to be explicit and the song itself is explicit too, so we think our model shouldn't have much trouble predicting it right. When we predict it with our model we get explicit, therefore we get it right again.

- Prediction for Rockstar, Da Baby: 0.7823029 , rounding, we get: 1 (explicit)

The last song, *Sunflower*, is a Pop song made by a more urban/hip-hop artist. The song is flagged as not explicit but the artist's music tends to be explicit (*see fig.6*). When predicted by our model we get an accurate prediction, which shows us that our predictor is quite robust and really has many variables that make him decide (many significant variables), not being biased to certain ones that could be seen as prejudices.

- Prediction for Sunflower, Post Malone 0.1427855 , rounding, we get: 0 (non explicit)

Conclusion

Machine learning opened a gate to a better understanding of the world. How do things work and how will things be are two questions that can be addressed through mathematically modeling a certain process. This kind of models are useful in a lot of different ways, and although not giving perfect predictions, they help us when taking decisions.

We all know when a song is explicit or not, we are humans and we understand our language. There are machines able to understand human language and therefore able to decide when someone us using strong language, but we have shown that there is also a correlation between the music and the language in a song, based on the song's characteristics.

At the end we have trained a model able to predict explicitness with a 95.9% test accuracy rate. We believe that this model could help music critics have a preliminary decision when rating a new song entering the market. We were also able to identify some errors produced in the predictions so we can easily assess future misclassifications.

A further and more useful study, could be the implementation of a dichotomizer that minimizes a certain risk function and the creation of a rejection class for ambiguous observations. By having a rejection class, we can save an observation for an expert to give a final decision on that song rather than misclassifying it. The risk function can also help censors avoid rating a explicit song as non explicit, which can lead to consumer dissatisfaction and bad examples for children.

Finally, the questions to be asked are: ¿Up to which point can we trust in the machine's decisions? ¿Will we rely on them when our job is at stake? ¿Who do we blame when something goes wrong? In the future, these will be the kind of questions which us, Data Scientists, will have to face when wanting to make a difference.

References

The data is available at: Billboard Hot weekly charts - dataset by kcmillersean

Spotify audio features obtainment: Get Audio Features for a Track

Some of the concepts used in the project not presented in the course: F1 score

Most of the mathematical models and procedures used in this project are based on the lectures of Machine Learning 1.

Annex

This project is free to use for educational purposes under citation.
For more information contact the developers via:
Abad Rocamora, Elias: *elias.abad@est.fib.upc.edu*
Fuentes i Oncins, Marc: *marc.fuentes.oncins@est.fib.upc.edu*
Martí Guiu, Alex: *alex.marti.guiu@est.fib.upc.edu*

fig.1

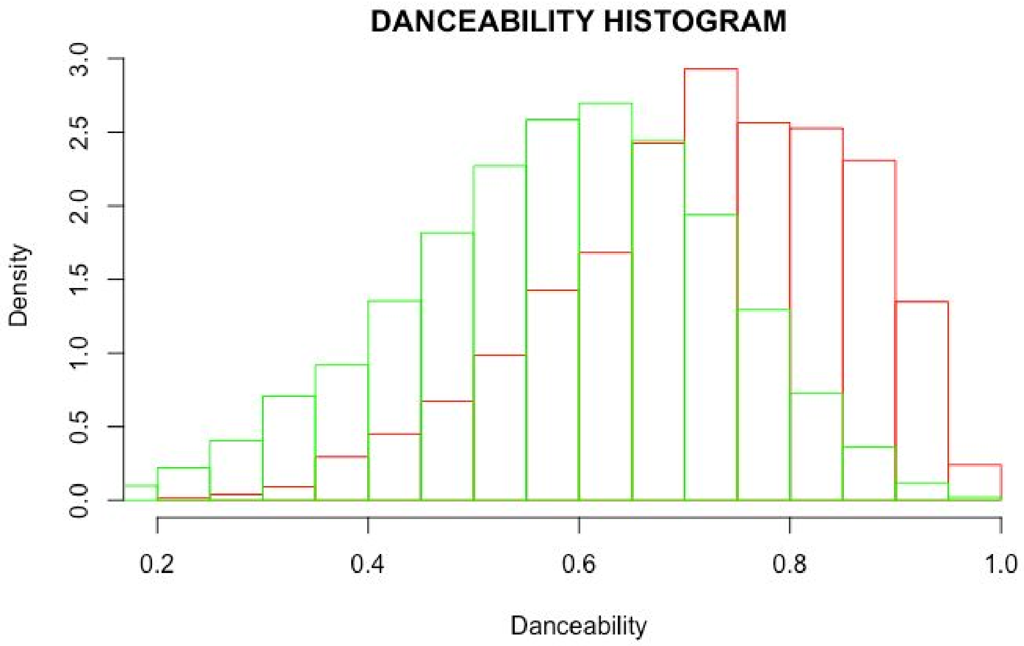
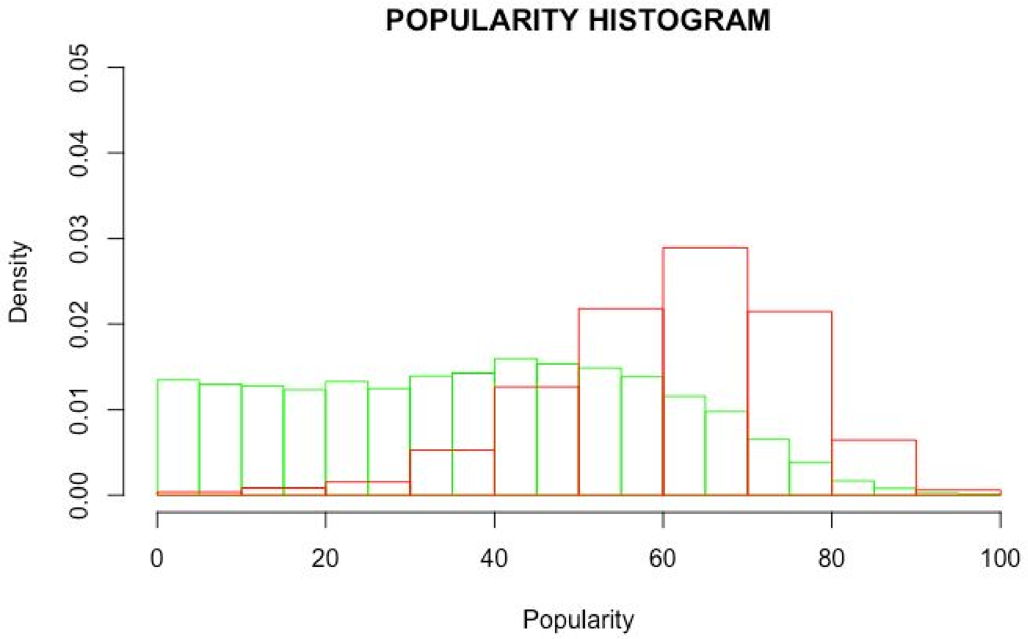


fig.2



POPULARITY HISTOGRAM
Density
0.00
0.01
0.02
0.03
0.04
0.05
0
20
40
60
80
100
Popularity

fig.3

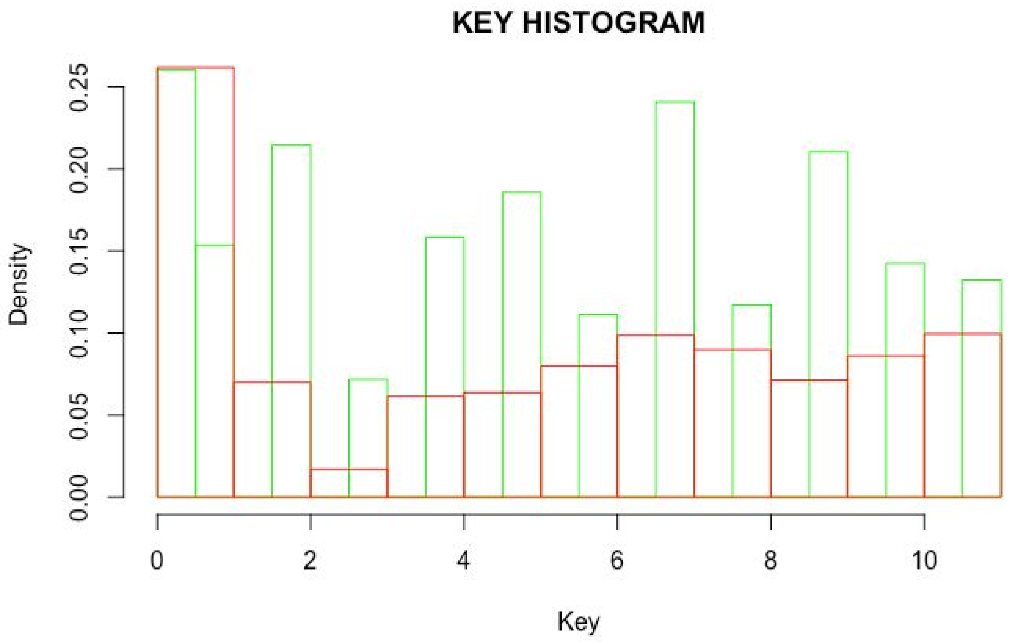

KEY HISTOGRAM
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
Key

fig.4

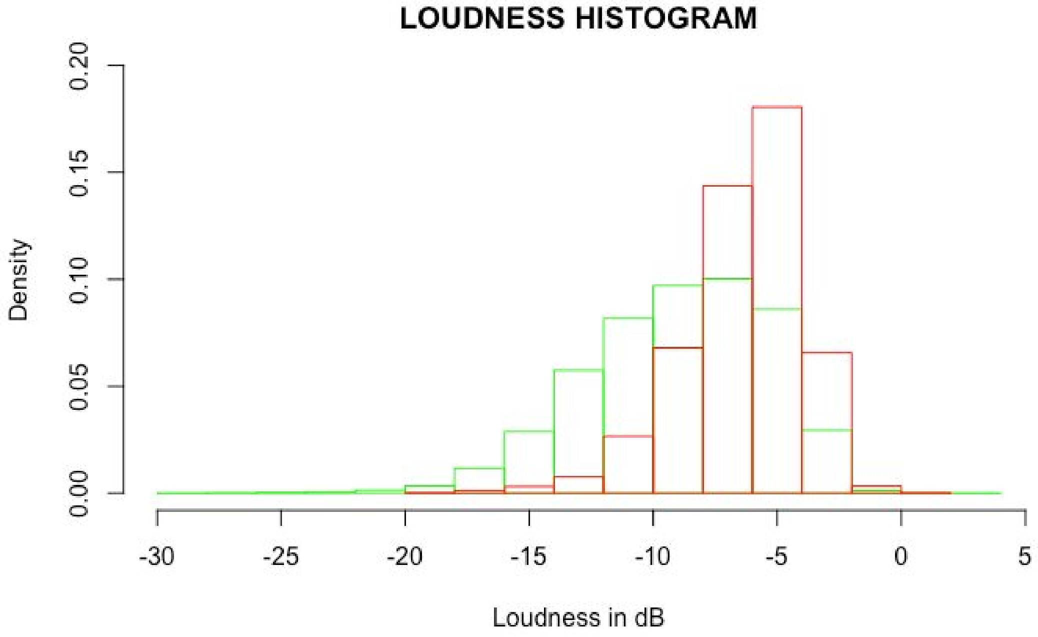

LOUDNESS HISTOGRAM
Density
0.00
0.05
0.10
0.15
0.20
-30
-25
-20
-15
-10
-5
0
5
Loudness in dB

fig.5

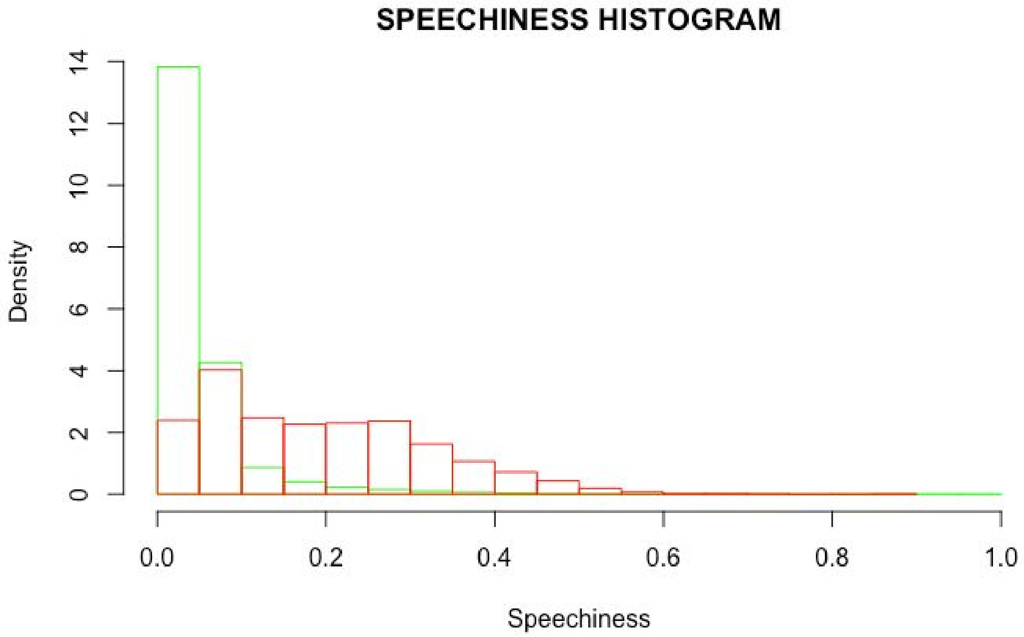
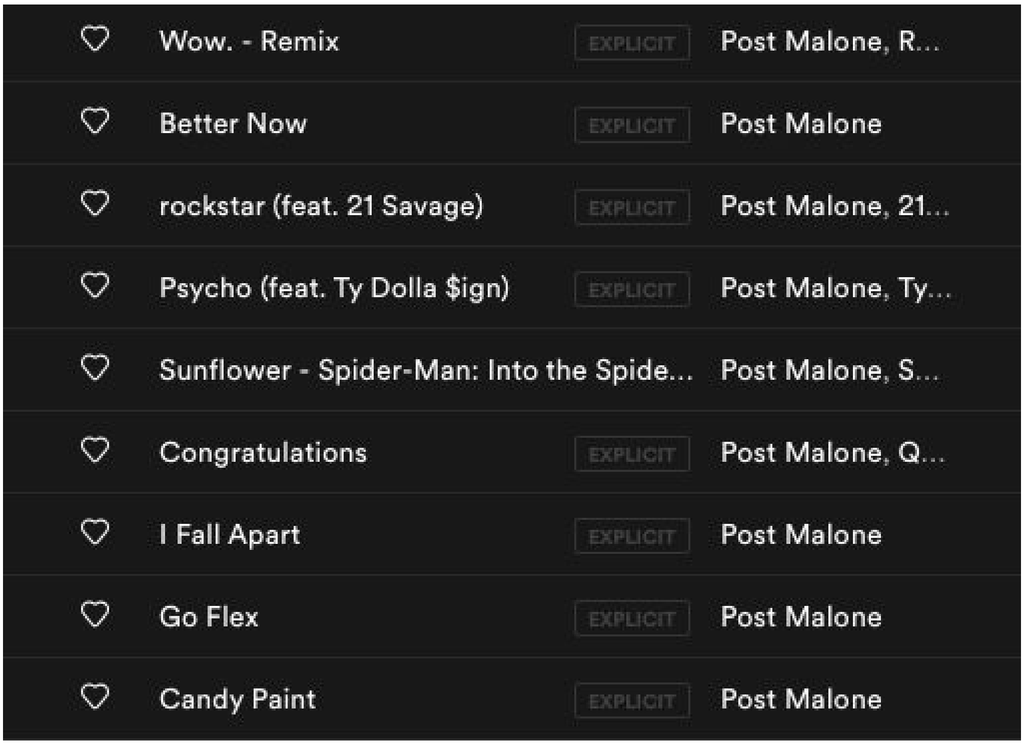


fig.6 *Post Malone's last album, where most songs are explicit.*



♡	Wow. - Remix	EXPLICIT	Post Malone, R...
♡	Better Now	EXPLICIT	Post Malone
♡	rockstar (feat. 21 Savage)	EXPLICIT	Post Malone, 21...
♡	Psycho (feat. Ty Dolla $ign)	EXPLICIT	Post Malone, Ty...
♡	Sunflower - Spider-Man: Into the Spide...		Post Malone, S...
♡	Congratulations	EXPLICIT	Post Malone, Q...
♡	I Fall Apart	EXPLICIT	Post Malone
♡	Go Flex	EXPLICIT	Post Malone
♡	Candy Paint	EXPLICIT	Post Malone