“Music was better in the old days” **said my parents... -** But I like it better ***today!!!***

Predicting year of release for songs from https://github.com/mdeff/fma

Svend - ckb742
Julius Jessen Terp - lft507

Goal

Determine whether there is a difference in music based on year of release, according to the features included, which were extracted using LiBROSA (https://librosa.github.io/librosa/)?

A little bit about the our data

Dataset: FMA: A Dataset For Music Analysis (~900gb) https://arxiv.org/abs/1612.01840

Using subset from: https://github.com/mdeff/fma

106574 songs, with 70294 including the date of release

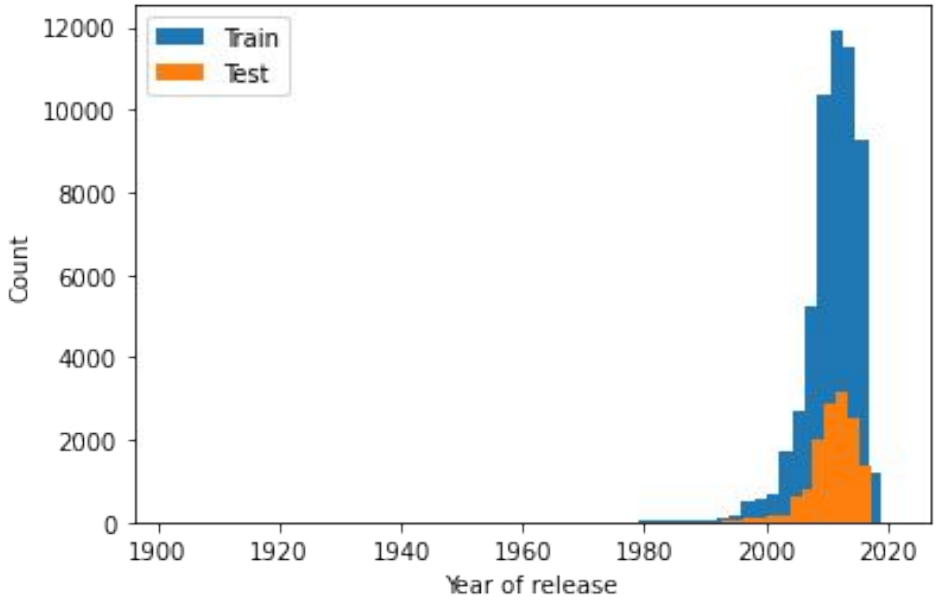
518 features extracted directly from the audio-files of the songs using LIBROSA

A glance at data

So… Unbalanced data?

Definately!

But we tried anyway!

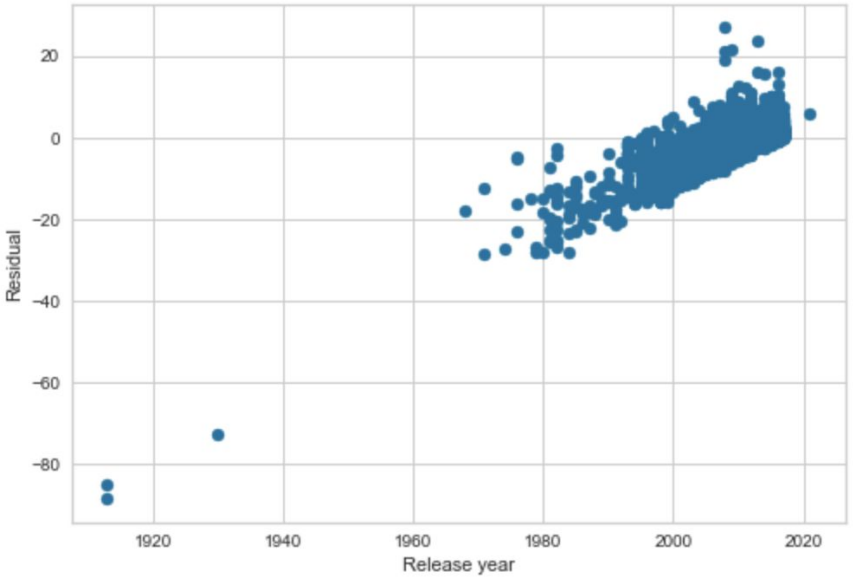


First try - Regression

LightGBM

MAE=1.656…

But the residuals don’t look that promising, at least for early years.



Multiclass Classification - first try

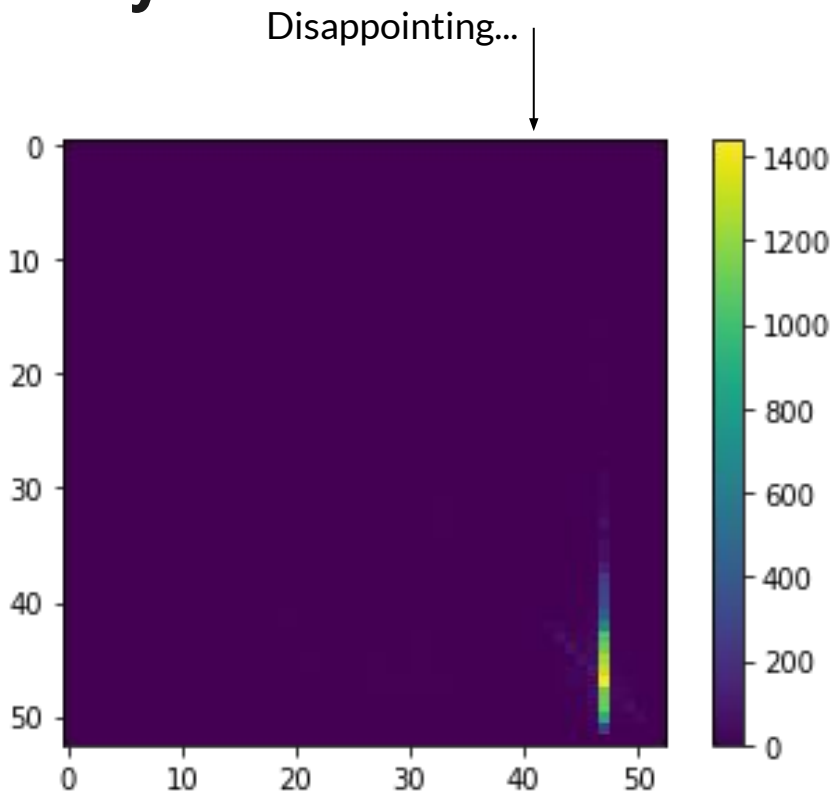
LightGBM: Multiclass

Each year is one class.

We optimized the BDT by boosting upto 100 and taking the best iteration, evaluated by logloss.

The best multi logloss was 2.9

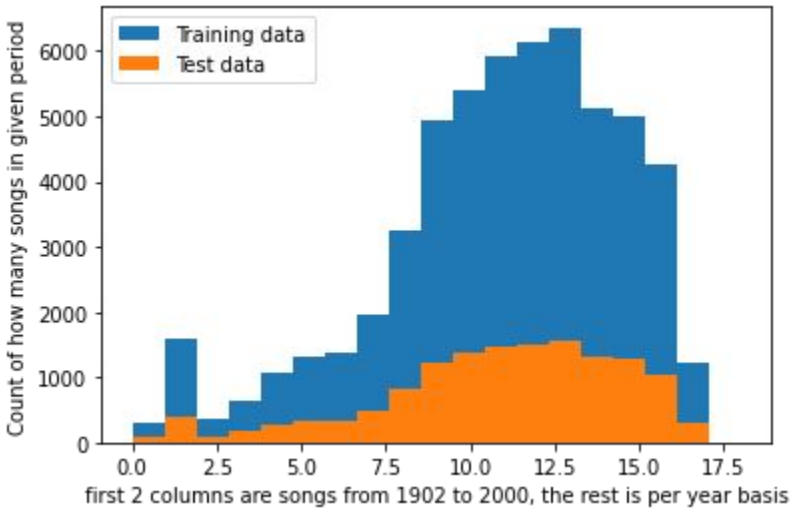
This plot doesn't show us much, as we know that we have only e.g. 1 song from 1902 (0 in the plot)



Restructuring data

Reducing number of classes

Done manually, by looking at the count in the different groups.



Multiclass Classification - second try

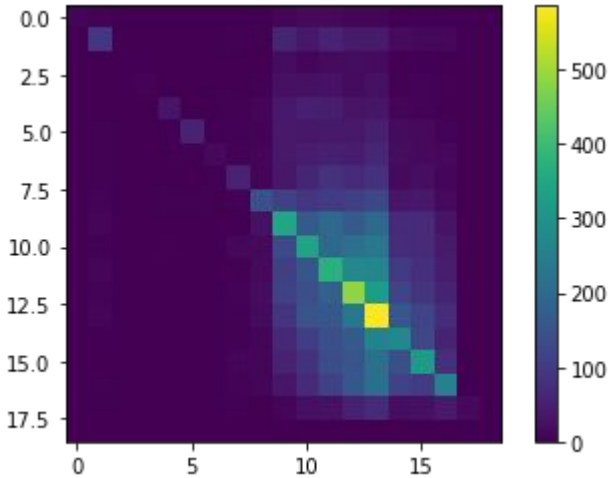
Using the before mentioned classes.

As expected, works better, as the data are distributed better between classes.

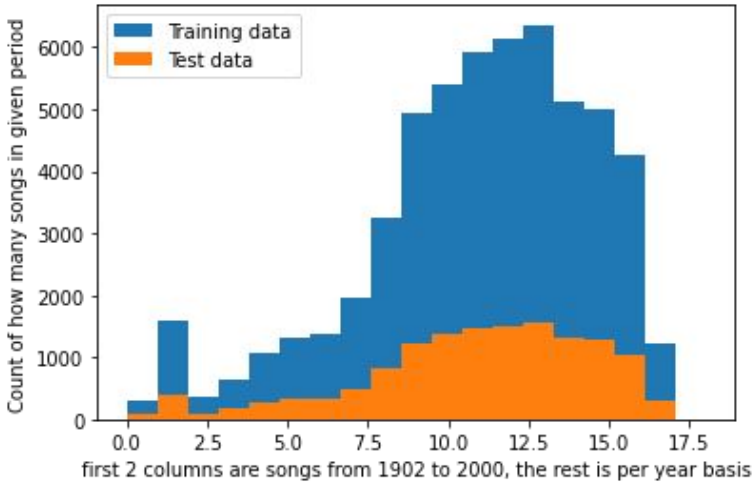
We optimized the BDT by boosting upto 100 and taking the best iteration, evaluated by logloss.

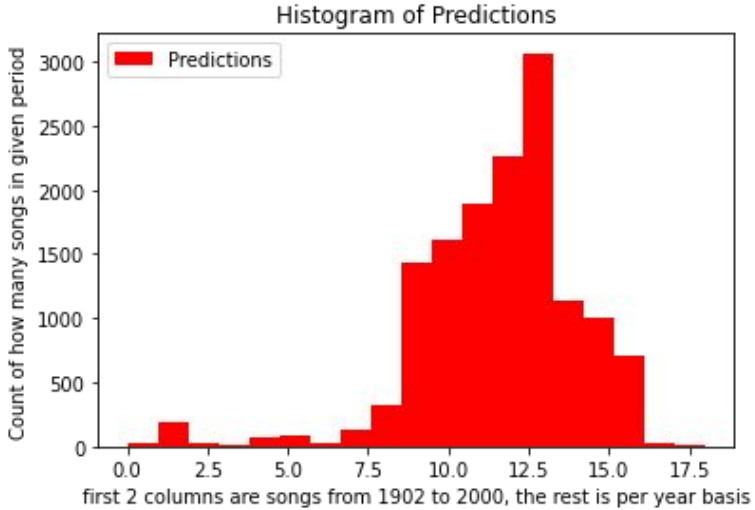
Best iteration had logloss of: 2.4

we tested how accurate it was, got a score of 23.8%



As histograms

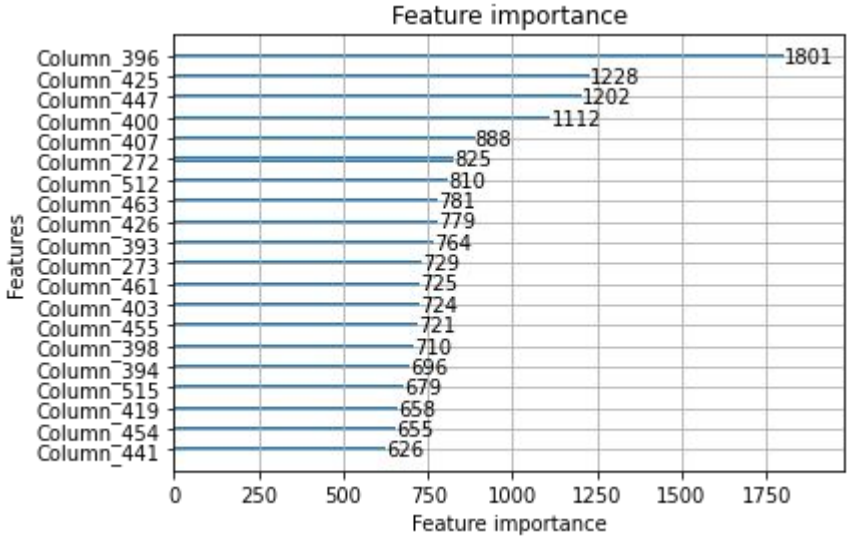




Most predictions are focused in the years between 2007 and 2017
just like the data

Feature importances

All these columns represent a feature of the audio spectra



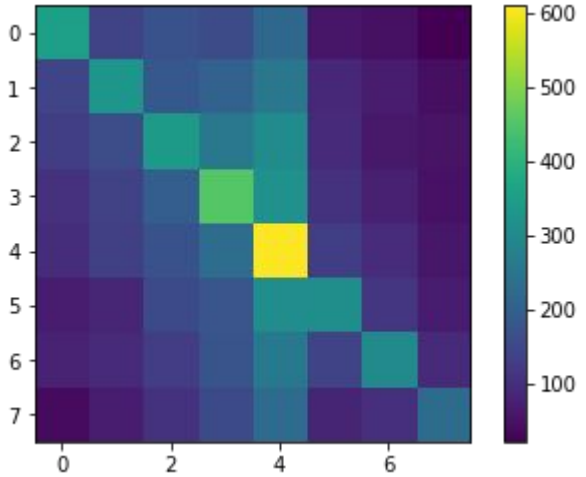
But what if songs are very similar? Let's take uncertainty into account

Let’s limit our time period: decade 2007-2017

Even with limiting our time period, our accuracy is still 23.8%

By taking uncertainty of 1 year, we get 52.8% correct

By making it 2 years. 70.7%

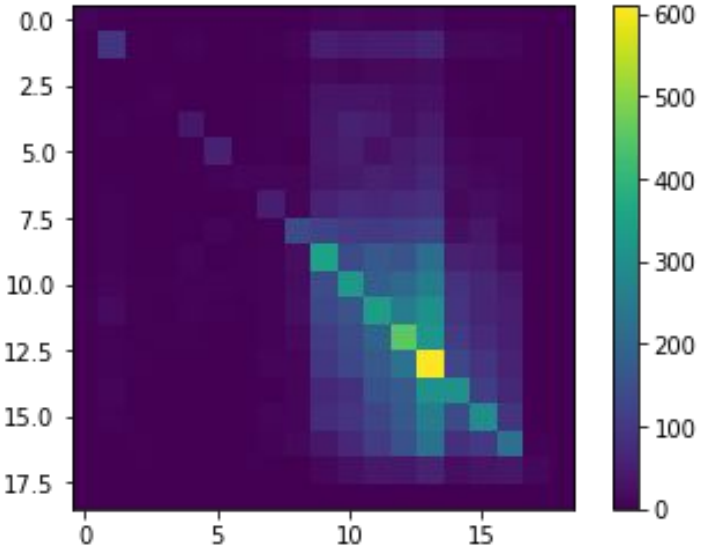


Uncertainty on all

No uncertainty; accuracy still 23.8%

1 year uncertain: 43.5%

2 year: 56% correct



Problems - and what to work on next

Didn't run clustering, to actually check if any correlations were to be found - could have run PCA, t-SNE etc.

Hadn't got the time to run e.g. Bayesian Optimization or RandomizedSearchCV, therefore HyperParameters were "guesses"

Didn't run a k-fold cross-validation.

Feature selection should most likely be done, to improve runtime, and thereby being able to include more data

GPU-implementation to be able to run more complex models, maybe NN

Unbalanced data…

Maybe:
- Focus on predicting the newer songs, where we have data available/more balanced data

Conclusions

Music changed - but it is hard to quantify how

It’s easier to predict release date, if you have a sufficient amount of data → Unbalanced data make it hard to make a model including everything.