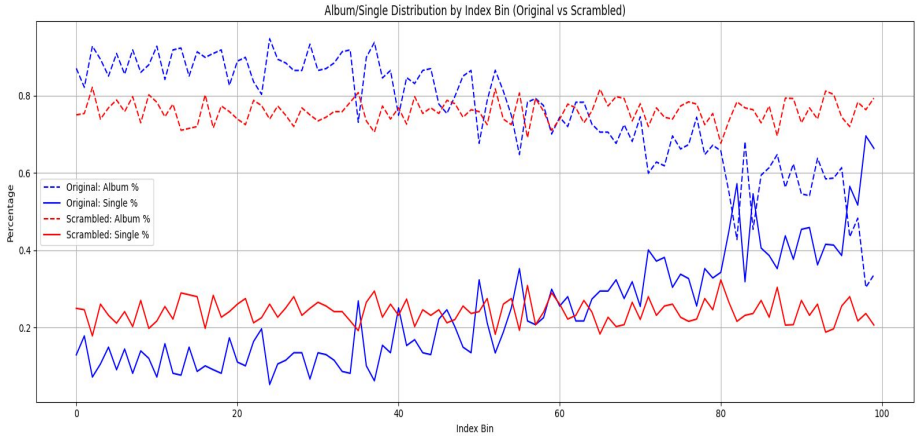
Machine Learning - EX2

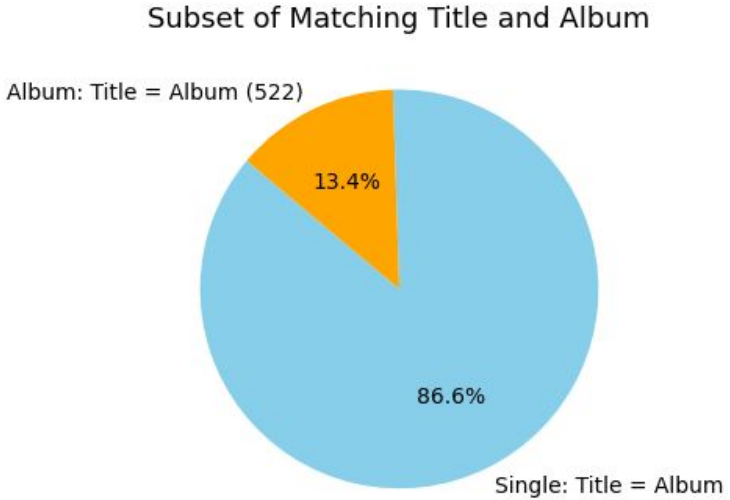
Insights & Conclusions

Ohad Livay, Tom Sapir

EDA

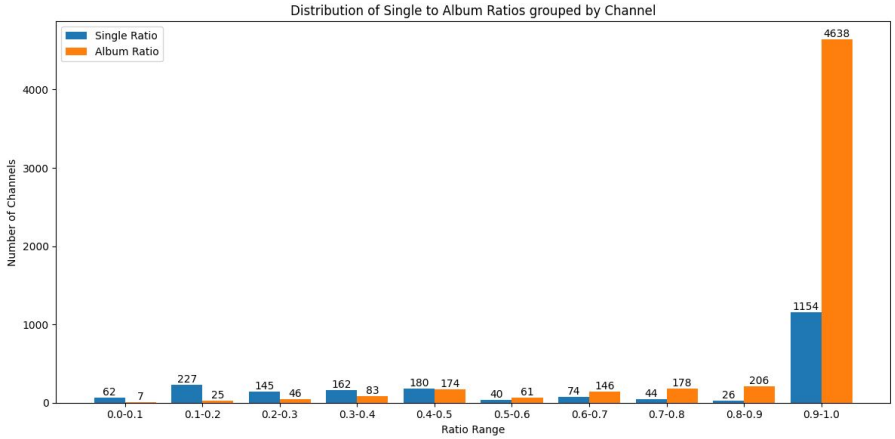


- Dataset order is biased
- Did not use this information
- Might have improved our scores
- Future inference will not have index, bad generalization

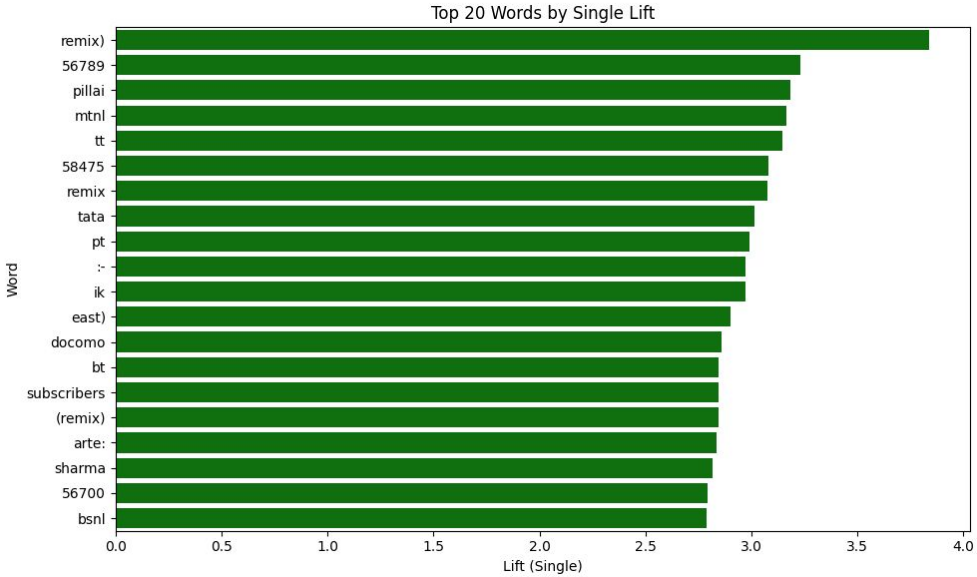


- Idea came from past experience
- *Golden Feature #1!*
- Demonstrates the strength of “String based features”
- pushed us to use bag of words

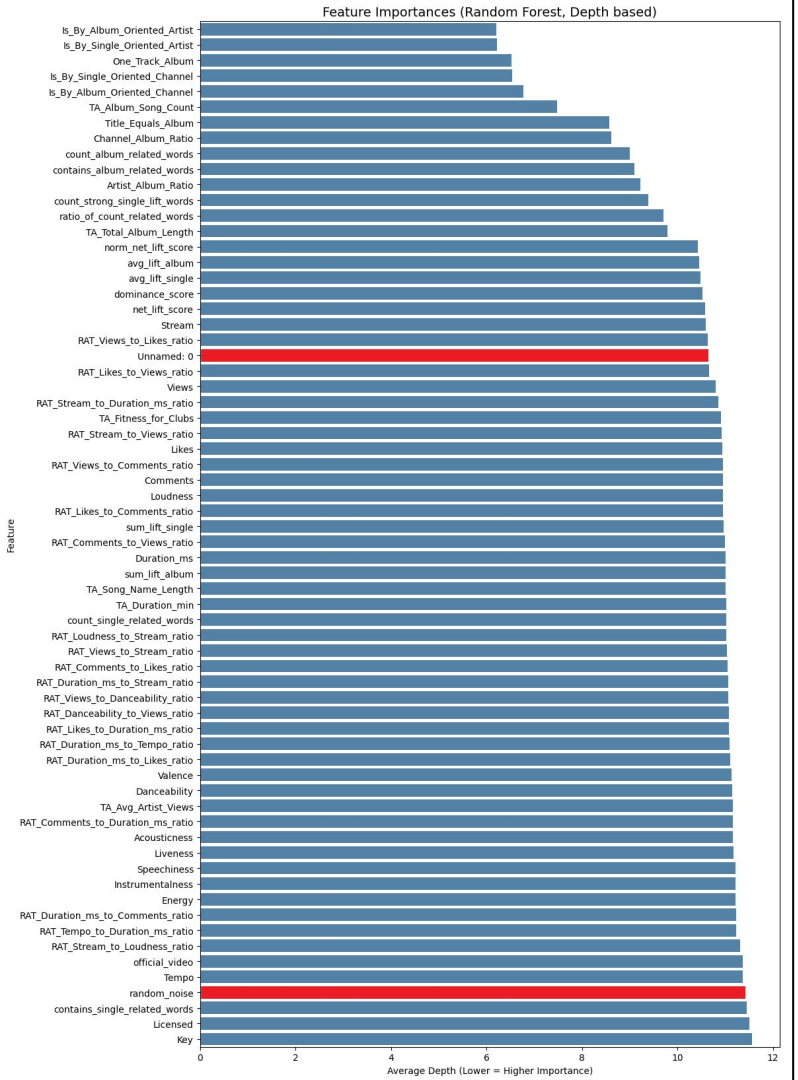
EDA



- most channels are purely album/single based.
- Even more so indicative of “Album songs”
- *Golden feature #2!*
-

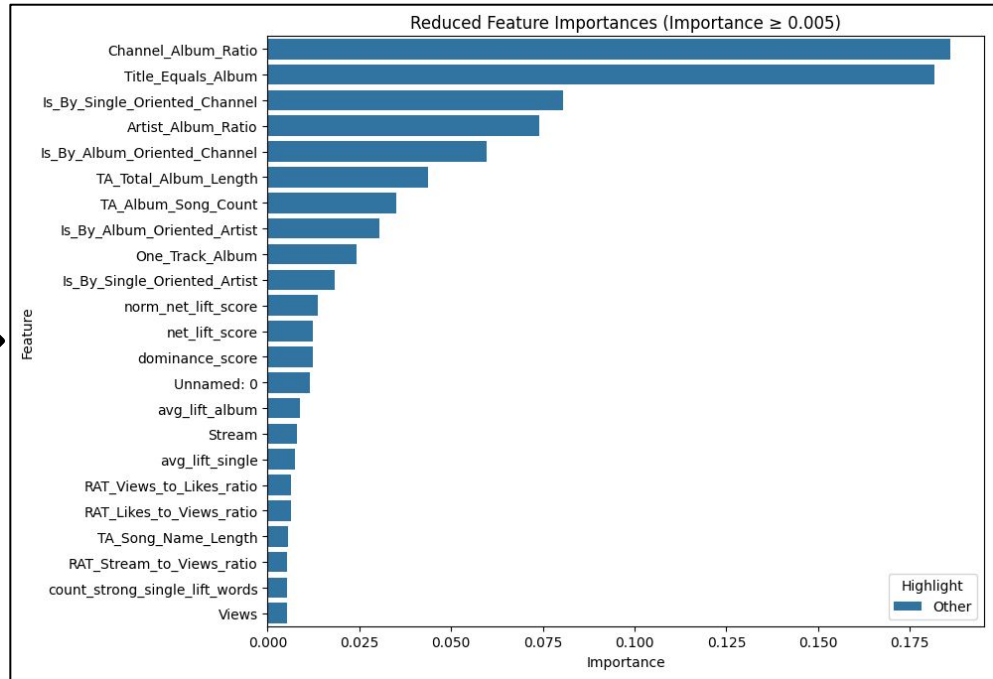


- "remix" lifts single
- Bag of words on string features. turn string -> numeric
- Lift metric cause imbalance
- Improve by parsing, tokenizing, trying things.. not in scope of this course
- words lifting in both single and album -> calculate a score



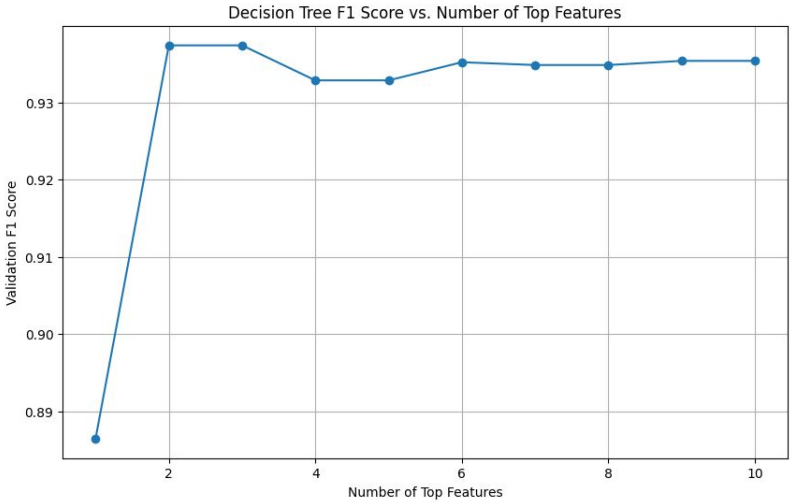
Feature Importance

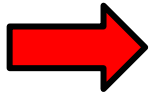
- importance for each feature using avg height in rf
- engineered features top the list
- add random noise for orientation of good vs bad features
- Unnamed: 0
- calculate correlation between all features and apply threshold to filter features.
- dynamically feed model top 1->n features

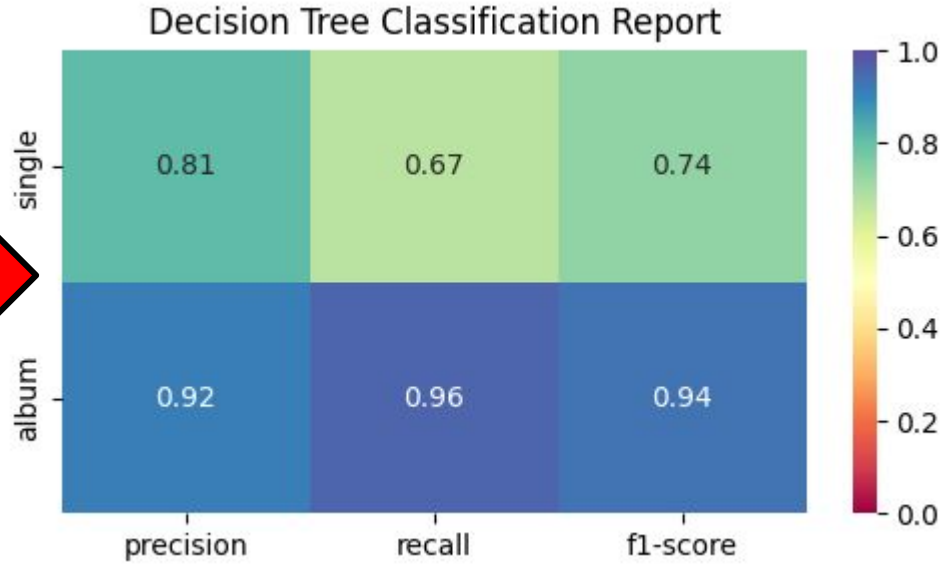


Best parameters found: {'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2}

Classifying album/single







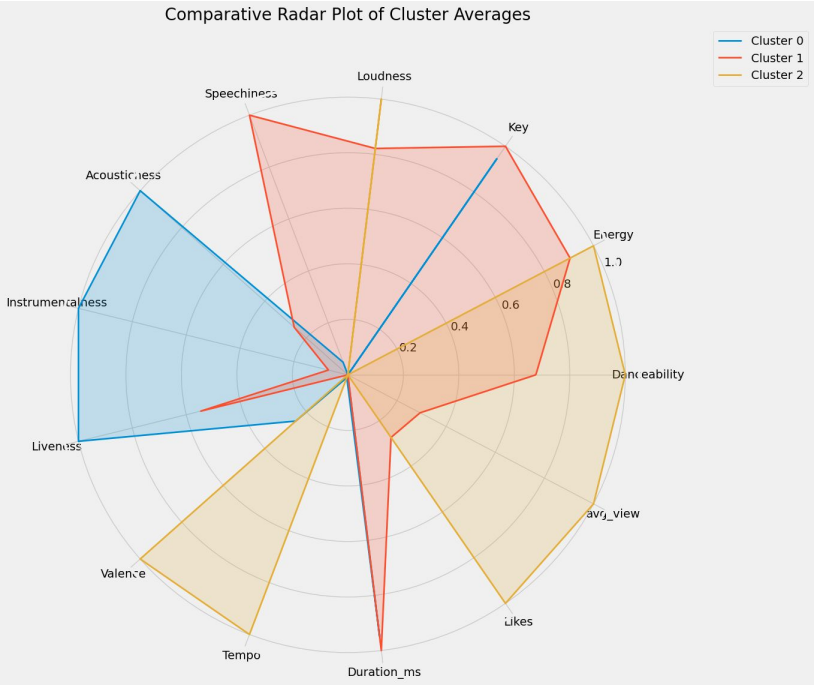
- tree models for non-normalized data
- choose top 1->n each time, get f1
- more features != better. even for trees based.
- top 2 features matter most

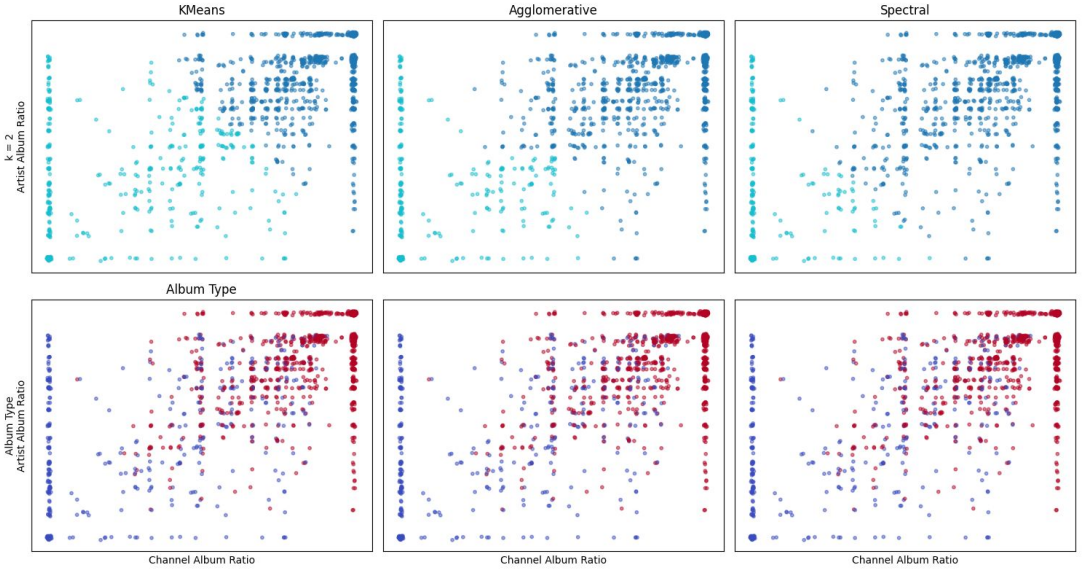
- model tuned for F1, not Accuracy (imbalance)
- Gap between the F1 scores of “Album” and “Single” - Each one in the role of “Positive”
- All models performed similarly

Clustering

- using only 2 features!!

grouping by Artist and clustering Artists





key to popularity: release short, loud, and upbeat songs