Predicting Song Popularity

Philippe Lessard, Petra Kumi, Camelia D. Brumar, Yanniode Peri-Okonny

Spotify Worldwide Daily Song Ranking

	# Position	A Track Name	A Artist	# Streams	A URL	Date
	Position on charts	Title of song	Name of musician or group	Number of streams		
	104.48 - 108.46 Count: 68,193 1 200	18597 unique values	Ed Sheeran 4% The Chainsmokers 2% Other (6626) 94%	1k 1.1m	21746 unique values	01/15/2017 Count: 71,7 31Dec16
1	1	Reggaetón Lento (Bailemos)	CNCO	19272	https://open.spotify.com/track/3AEZUABDXNtecAOSC1qTfo	2017-01
2	2	Chantaje	Shakira	19270	https://open.spotify.com/track/6mICuAdrwEjh6Y6lroV2Kg	2017-01
3	3	Otra Vez (feat. J Balvin)	Zion & Lennox	15761	https://open.spotify.com/track/3QwBODjSEzelZyVjxPOHdq	2017-01
4	4	Vente Pa' Ca	Ricky Martin	14954	https://open.spotify.com/track/7DM4BPaS7uofFul3ywMe46	2017-01
5	5	Safari	J Balvin	14269	https://open.spotify.com/track/6rQSrBHf7HlZjtcMZ4S4bO	2017-01
6	6	La Bicicleta	Carlos Vives	12843	https://open.spotify.com/track/0sXvAOmXgjR2QUqLK1MltU	2017-01

Spotify Song Features

track_name	track_id	popularity	acousticness	danceability	duration_ms
130254 unique values	153685 unique values	0 100	0 1	0.06 0.99	15.5k
Stiffelio, Act III: Ei fugge! … Lina, pensai che un angelo … Oh gioia inesprimibile	7EsKYeHtTc4H4xWiTqSVZA	21	0.986	0.313	
Madama Butterfly / Act 1: ... E soffitto e pareti	7MfmRBvqaW0I6UTxXnad8p	18	0.972	0.36	
Turandot / Act 2: Gloria, gloria, o vincitore	7pBo1GDhIysyUMFXiDVoON	10	0.935	0.168	
Rigoletto, Act IV: Venti scudi hai tu detto?	02mvYZX5aKNzdqEo6jF20m	17	0.961	0.25	
Don Carlo / Act 4: "Ella giammai m'amò!"	03TW0jwGMGhUabAjOpB1T9	19	0.985	0.142	
D'amor sull'ali rosee	0G75cCcf6vBSnMFFkVW9pq	20	0.99	0.211	
Waxman : Carmen Fantasie	10gPtjlpTS9Uq6EUQuGljt	13	0.98	0.341	

Project Goal

- Predict song popularity in a given country based on song features
- Merge datasets to connect features to regional chart position
- Regression might be hard to get good accuracy
- Classify into different categories:
  - Top 10 (1-10)
  - Top 50 (11-50)
  - Top 100 (51-100)
  - Top 150 (101-150)
  - Top 200 (151-200)

Merging Datasets

1. Drop irrelevant columns (track name, artist name, date)
2. Flatten genre column
3. Keep each song's peak position in each country's chart, and drop the rest
    - As songs move up and down the charts, their position changes
    - Features stay the same
4. Convert track_id to URL
5. Merge features dataset into rankings dataset using URL as a key
    - Equivalent of a SQL Natural Join
6. 3.4 million rows → 12,000
7. Split into 2,000 row safe

Feature Engineering - Categorical Data

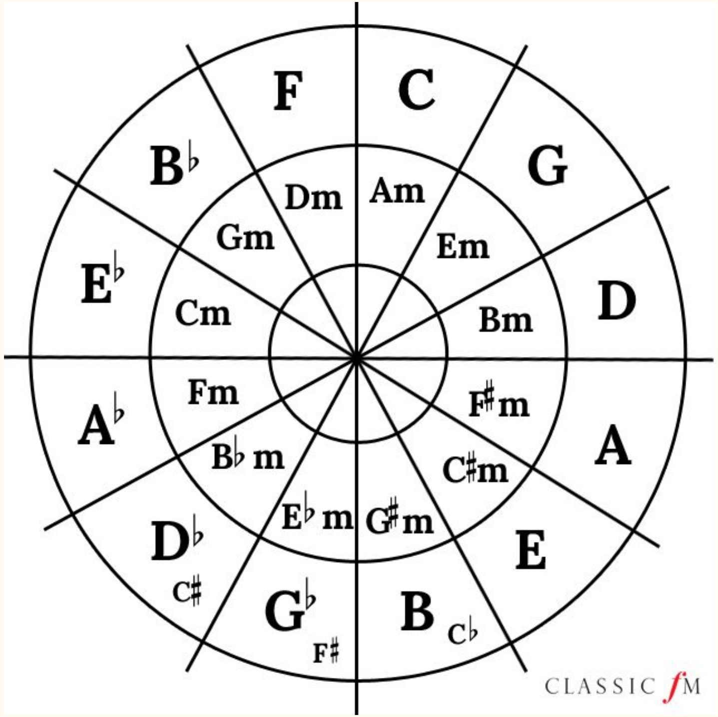
genre
Movie
Reggae
Jazz
Dance
Pop
Comedy
Reggaeton
Opera
Blues
Alternative
Anime
Children's Music
Rock
Folk
Indie
World

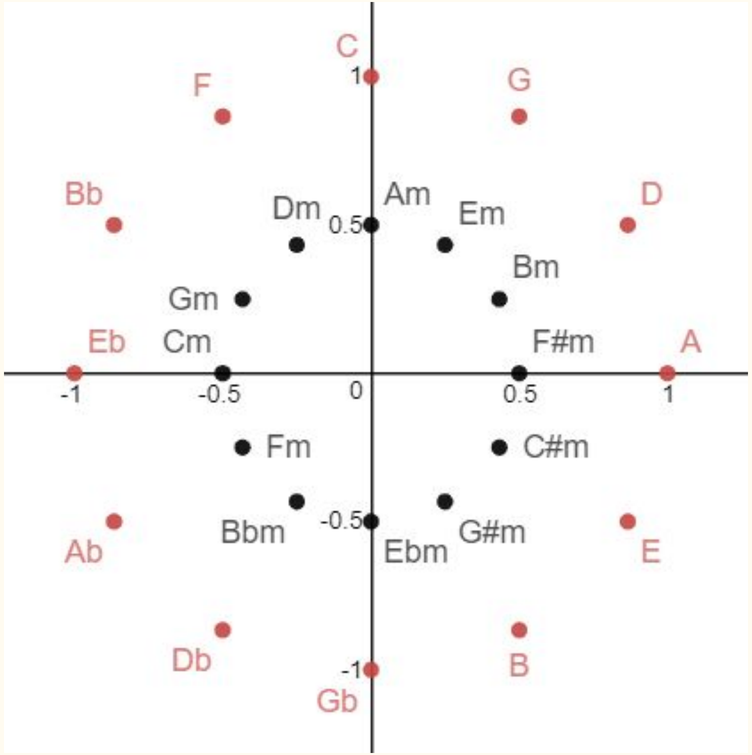
Blues	Dance	Pop	Electronic	R&B
0	1	1	0	0
0	0	1	0	0
0	1	1	0	0
0	1	0	0	1
0	0	0	0	0
0	1	0	0	0
0	0	0	0	0
1	0	1	0	1
0	0	1	0	0

Region
ec
ec
ec
ec
ec
ec
ec
ec
ec
ec

Region_cl	Region_ec	Region_ee	Region_es	Region_fi
0	0	0	0	0
0	0	0	0	0
0	1	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Feature Engineering - Key

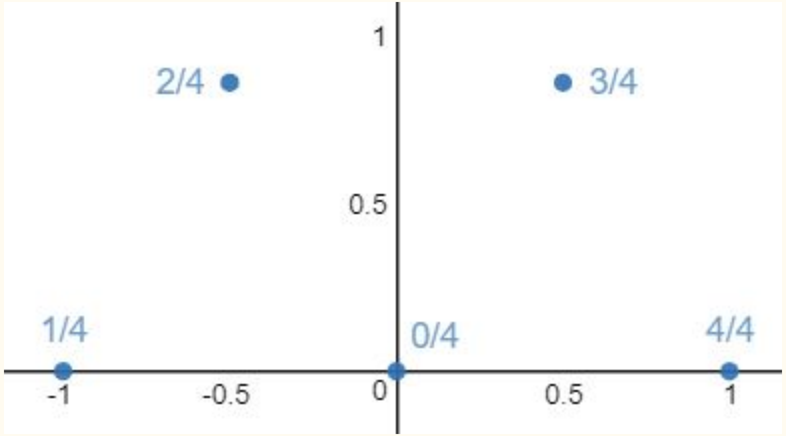




https://assets.classicfm.com/2018/13/circle-of-fifths--1523016231.jpg

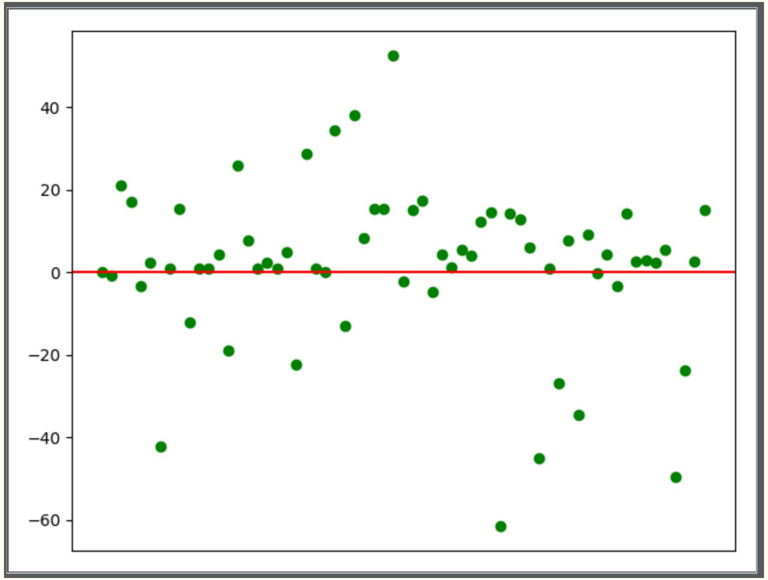
Feature Engineering - Time Signature

- 0/4
- 1/4
- 2/4
- 3/4
- 4/4



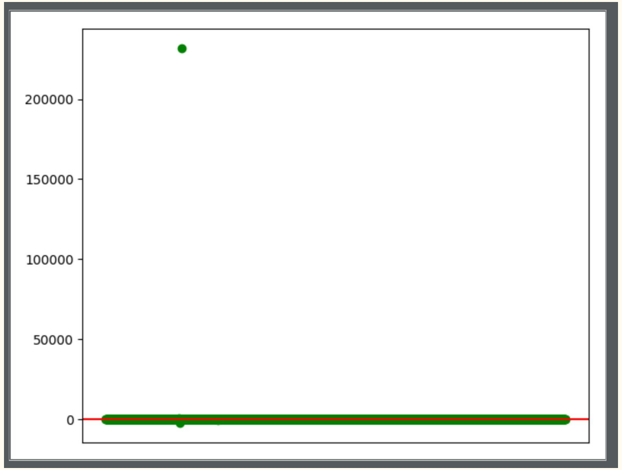
Lasso & Ridge Regression

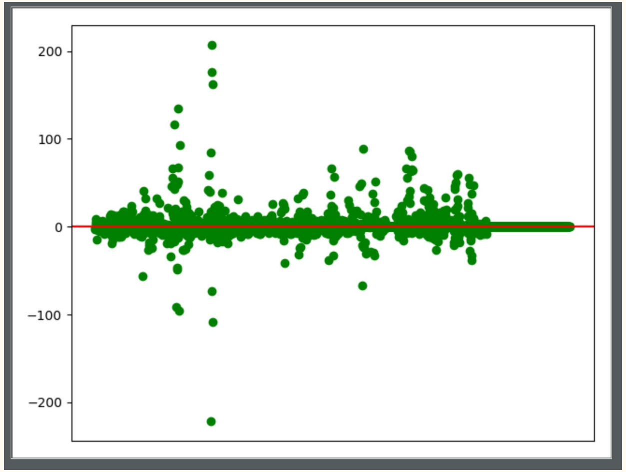
- Lasso: too strict - none of the predictors were deemed influential enough
- Generalized Cross-Validation used to determine alpha
- Most influential predictors:
    - Region_ee (Estonia)
    - Soundtrack
    - Region_sk (Slovakia)



Lasso & Ridge Regression

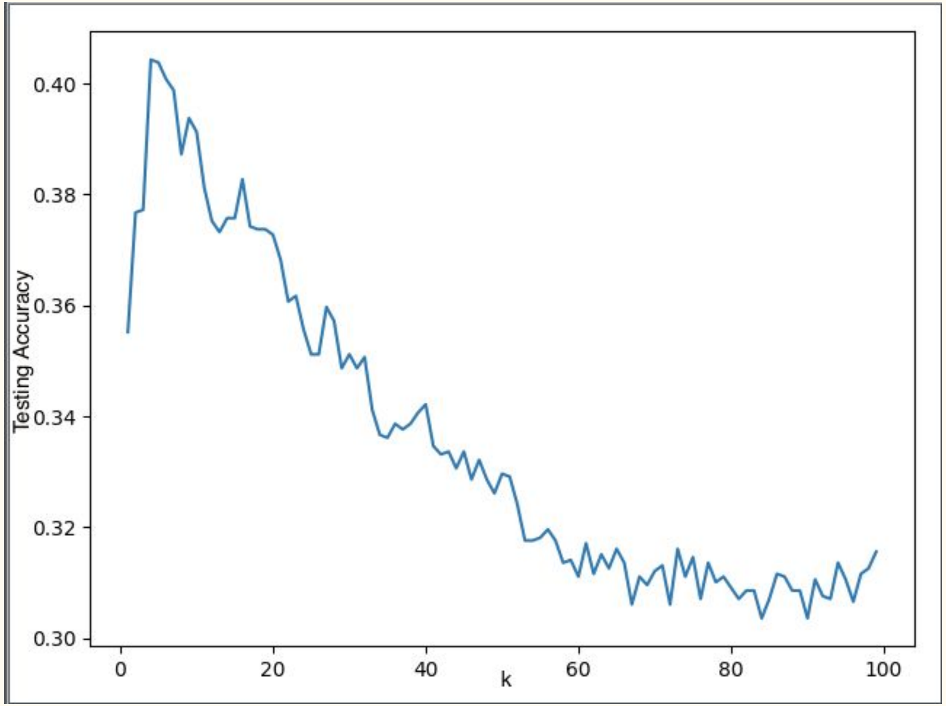
- Ridge Regression with interaction terms
- Most influential parameter:
  - instrumentalness * movie
  - instrumentalness * Reggaeton
  - instrumentalness *Reggae
- MSE = too high
- Figure 1: Results of ridge regression (with interaction terms).
- Figure 2: Results of ridge regression - all predictors except for the three most influential ones.





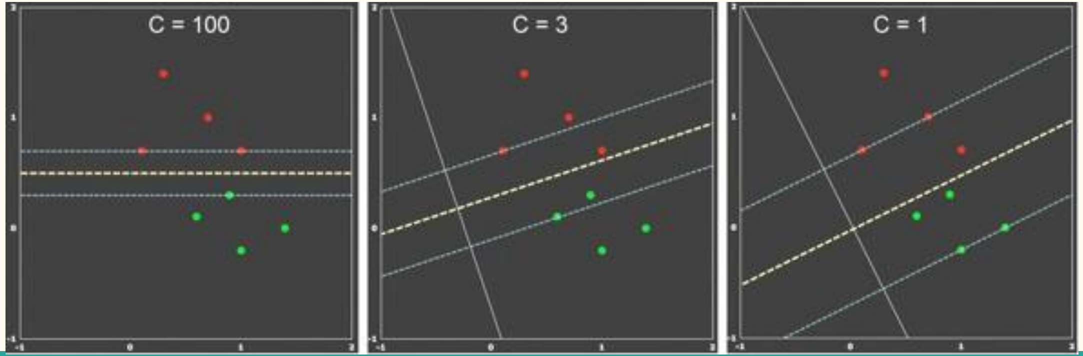
KNN and KNN + PCA

- KNN: no strict assumptions about the data
- K-fold cross validation
- Best accuracy:
    - Only KNN:
        - $K = 10$,
        - $Accuracy = 0.3852$
    - KNN + PCA
        - 29 principal components,
        - $K = 7$
        - $Accuracy = 0.3859$



SVM

- Types of Kernel: Linear, RBF, Polynomial, or Sigmoid
- To determine the degree to use for a polynomial kernel
  - tested values between 3 and 10
  - degree of 3 had the best accuracy
- Set the C-parameter: tested values of 1e-5 to 10
- Best accuracy: RBF kernel with C=10
- Larger values of C tested manually
  - C=10,000, accuracy = 54.2%



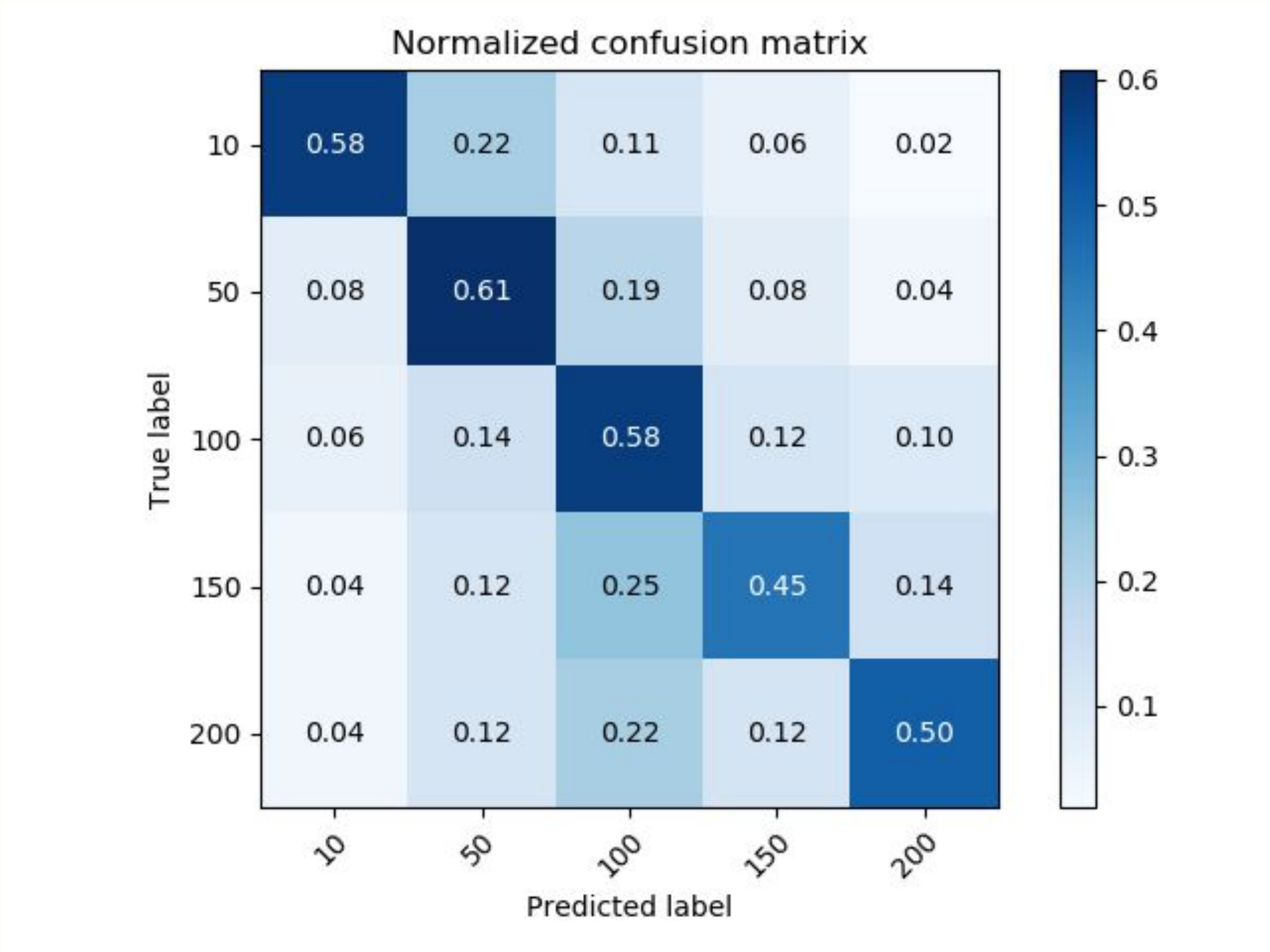
SVM

- Training: sample sizes usually different → bias in the training
- Solution: up-sampled the smaller class to the size of the larger class by bootstrapping
    - Samples were balanced
    - No loss of data by down-sampling
- Chose SVM model with higher testing accuracy

SVM

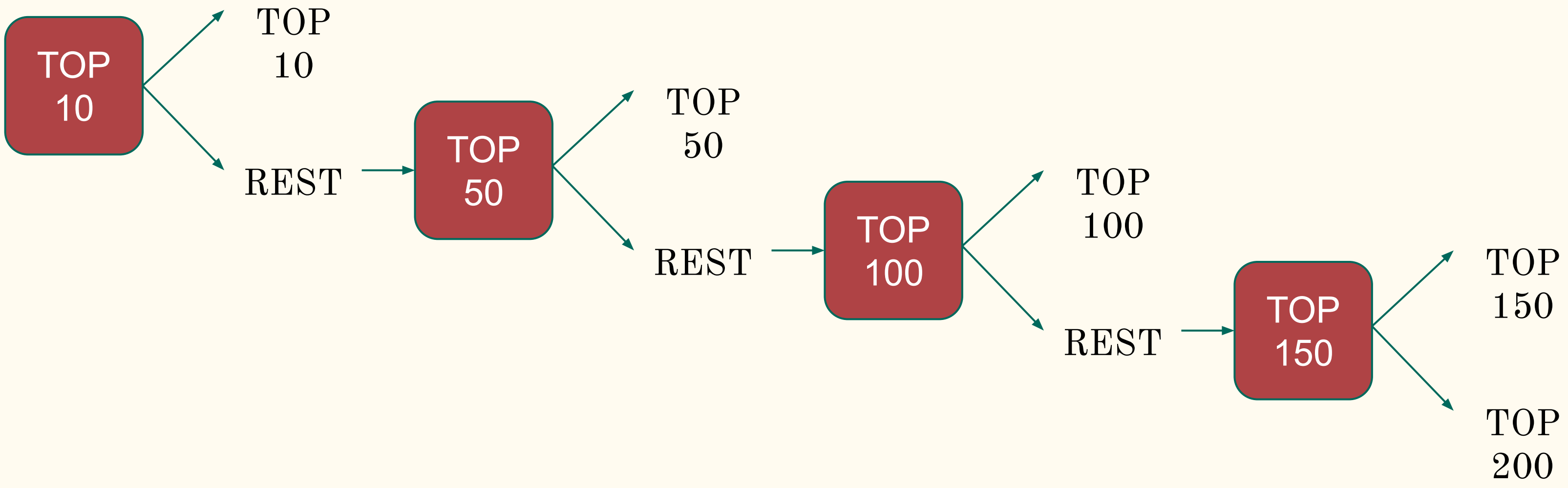
Larger values of C + RBF

- C=10,000,
- accuracy = 54.2%

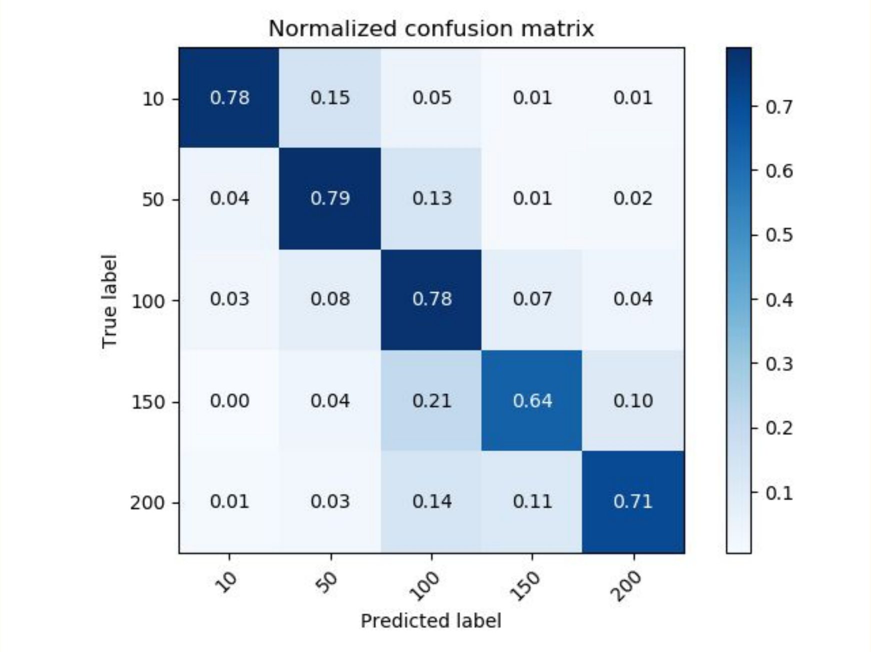


SVM

- 2 different **strategies** for using SVM:
  a. Multi-class SVM algorithm
  b. Layered set of one-vs-all binary classifiers
- Trained and cross-validated 4 different SVM models



Random Forests



- Expected not to overfit
- Used K-Fold Cross-Validation
- Tested for:
  - 70-150 trees with increments of 1
  - 100-500 trees with increments of 100
  - 3-12 folds
  - Max depth 20-40

Highest Accuracy For:

- Number of trees = 100
- Maximum depth of each tree = 30
- Number of folds = 6
- Accuracy score = 0.7405

Summary of the Final Approach

1. Cross validating to find the optimal number of hyperparameters for the models
2. Tested their accuracy on sets of testing data with the goal of finding the most successful model
3. Best model: a single multi-class Random Forest with 100 trees, 6 folds, and 30 as our maximum depth
   - accuracy = 0.7405
4. Followed by a single multi-class SVM, and KNN with 10 neighbors and 29 PCA components.
5. Chosen Random Forest (no overfitting)
6. Proceeded to test it against the data we had put in the safe.

~70%

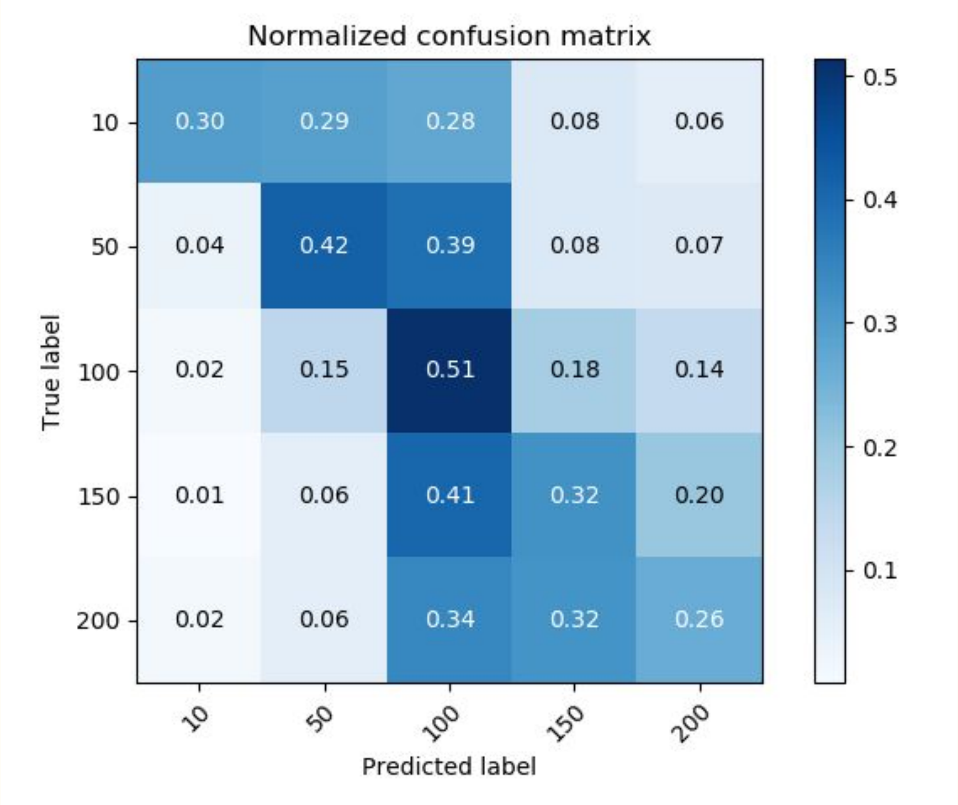
Expected Accuracy

37.4%

Final test accuracy

Summary of the Final Results

- 30% True Positives for predicted top 10 songs
- 51% True Positives on top 100 data
- 26% of the data points rightfully classified as top 200.

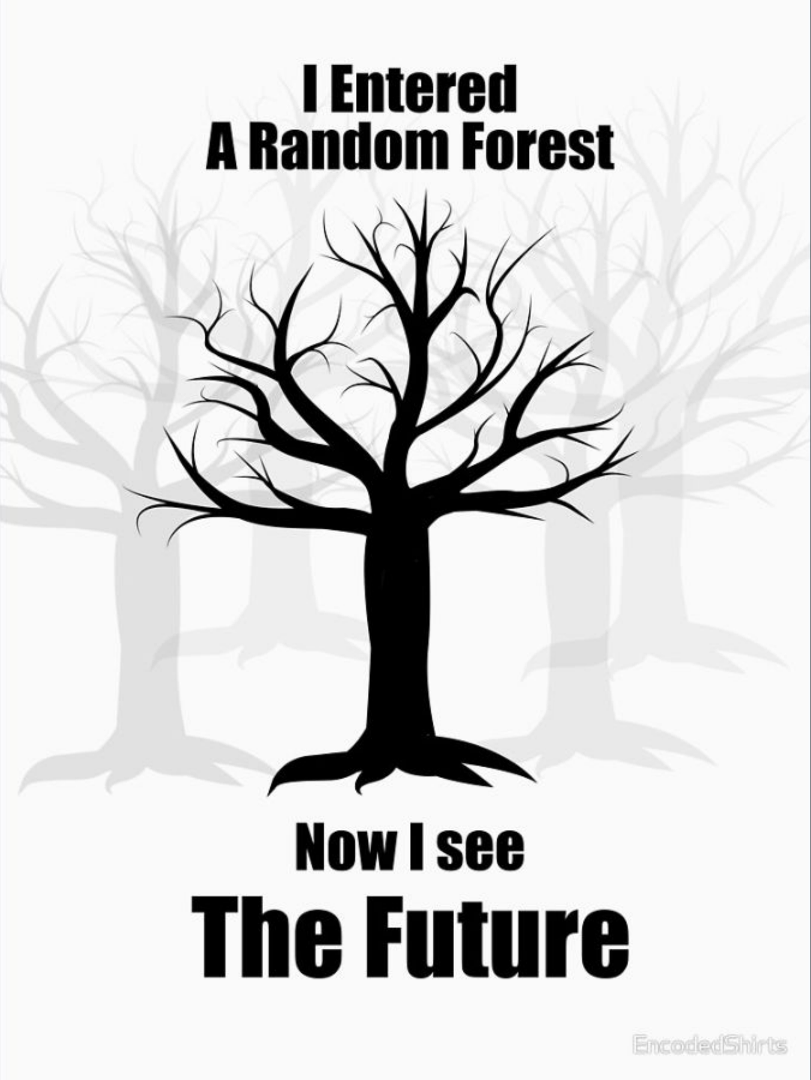


Summary of the Final Results

- Our model does well in predicting the general trends of a song
- Possible reasons behind low model accuracy
  - Accidental data snooping
  - May have overlooked/misunderstood how some functions are used
  - Chose final model based only only average accuracy, not true/false positives or negatives
- May have picked a more biased model, much more accurate on true negatives than it is on true positives.

Conclusions

- Lasso, Ridge Classification, KNN, KNN + PCA, Random Forest, and SVM.
- Merged 2 datasets
- Did feature engineering to make the dataset useable for our purposes
- Stored some of the data in “the safe” to only use for testing of the final model.



Conclusions

- Used k-fold cross-validation and bootstrapping.
- Fine-tuned our models by cross validating on different hyperparameters.
- Most accurate model: Random Forest with 100 trees, of 30 maximum features each, and 6 folds for k-fold cross-validation.
- When running it on our safe data → lower accuracy than expected.
- Suspecting accidental data snooping or not have examined model accuracy scores in more depth.