Music Genre Classification

Group 7 - SENG 474 Project Presentation - Spring 2022

Jeet Ajmani

Jagjeet Singh

Kenil Shah

Ella Kuypers

Can we classify a song’s genre using only its Spotify metadata?

Our project

- Genre classification using only music metadata
  - Avoids using computationally intensive algorithms
  - Might uncover trends not visible when using standard audio machine learning algorithms (beat tracking, tempo, amplitude, etc.)
  - Helpful in cases where access to audio file might not be a possibility

Track Title
Primary Artists
Featured Artists
Publisher
Genre
Subgenre
Composer
Producers
Additional Contributors
WAV
Year of Recording
Year of Composition

Our project

- If you have the metadata, don’t you already have the genre?
  - We are interested in applying machine learning to music but wanted to see if it was possible without audio and only basic features
  - This project is scoped to genre classification but we were curious to see if we could predict other labels, like maybe even guess the title (Think of AI movie script generators)
  - Could have applications that may not seem obvious

The Dataset

- Spotify Daily Charts over 3 years
    - Data is mostly cleaned
    - Contains metadata for over 50,000 songs
    - Includes huge number of useful features to do genre classification
    - Example: danceability, energy, loudness, speechiness, acousticness, duration

Example of the Dataset

Title	Artist	Genre	…	Energy	Tempo	Duration (ms)
adan y eva	Paulo Londra	argentine hip hop	…	0.709	171.993	258639
it wont kill ya	The Chainsmokers - Louane	dance pop	…	0.538	170.138	217613
arrows	Foo Fighters	alternative metal	…	0.917	121.958	266187
…	…	…	…	…	…	…
scooby doo pa pa	Dj Kass	chilean hardcore	…	0.754	120.939	145972
positions	Ariana Grande	dance pop	…	0.802	144.015	172325

Dataset Problems

- Double-edged sword
  - Dataset cannot be used in its raw form, most of the difficulty is in preprocessing and feature selection
  - Many features require preprocessing such as scaling (numerical) or one-hot encoding (Strings)
  - Not all features may be helpful with classification, need careful feature selection

- Duplicate song entries
  - Dataset contained multiple rows for each region where it charted, including Globally

- Missing values, even in the genre column
  - Dropped all rows with any “crucial” missing values right away

Initial Experiments

- All algorithms were failing to classify when given the raw dataset

	Experiment 1: *No Numerical Preprocessing*		**Experiment 2:** *Minmax Preprocessing*		**Experiment 3:** *Standard Scaler Preprocessing*	
Algorithm Name	**Training Error**	**Test Error**	**Training Error**	**Test Error**	**Training Error**	**Test Error**
Decision Tree	0.0	0.337	0.0	0.315	0.0	0.335
Random Forest	0.139	0.397	0.0	0.354	0.0	0.364
K Nearest Neighbors	0.445	0.631	0.394	0.535	0.388	0.561
Linear SVM	Did not converge	Did not converge	0.512	0.521	0.516	0.530
Logistic Regression (sag)	0.582	0.402	0.505	0.513	0.502	0.519
Logistic Regression (lbfgs)	Did not converge	Did not converge	0.505	0.513	0.502	0.519

Overfitting

Noisy Dataset

Experiment #1

Determining the best machine learning model for our problem

Sub-tasks:

- Data preprocessing
- Feature selection
- Model performance assessment

Preprocessing: Feature Matrix (X)

- Non-numerical features such as Artist
  - Utilized one-hot encoding for Artists based on Song Title
  - Too many new columns resulted in sparse features and poor performance from Decision Trees

Song	ArtistA	ArtistB	ArtistC
Song1	0	1	0
Song2	1	0	0
…	…	…	…
SongN	1	0	0

- Standard Scaling applied to numerical features
  - Significantly improved the performance of Linear SVM

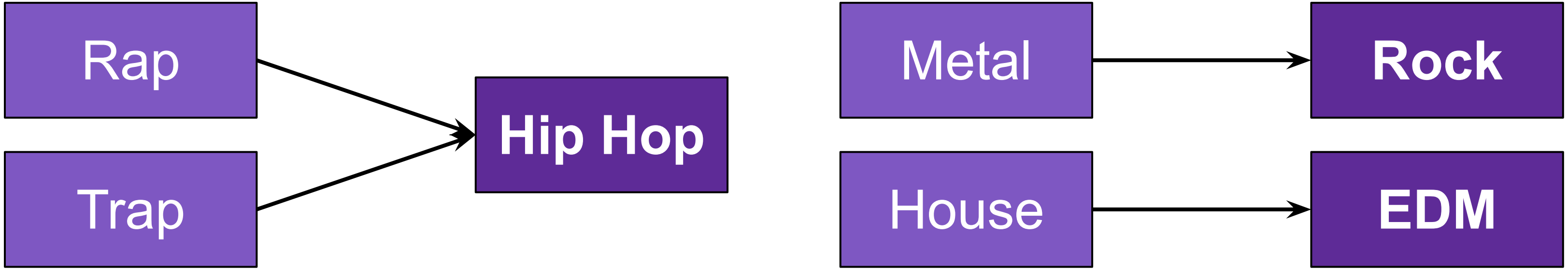
- Features extracted from lyrics only existed for English songs
  - Many rows containing non-English songs had missing data
  - Experimented with (1) dropping rows containing missing data, (2) initializing missing data to 0, (3) ignoring feature columns with missing data

Preprocessing: Target Labels (y)

- Dataset contained hyper-specific genres (Canadian Hip Hop, Dance Pop, etc.)
  - Also contained a column of parent genres (Hip Hop, Pop) that narrowed the target label down

- Early experiments revealed slight classification mistakes
  - Predicting “Hip Hop” instead of “Rap”
  - Predicting “Rock” instead of “Metal”

- Decided to group some genres into even larger groups



Without Feature Selection

- Used an arbitrary set of basic features available for all songs
- Performance with stratified k-fold:

	Decision Trees	Random Forests	K-nearest neighbours	Logistic Regression	SVM
Train Error	0.333	0.294	0.415	0.316	0.215
Test Error	0.345	0.301	0.442	0.298	0.192

- SVM delivers best results due to the high number of one-hot encoded features

Tree-based Feature Selection

- What is tree-based feature selection?
- Performance with stratified k-fold:

	Decision Trees	Random Forests	K-nearest neighbours	Logistic Regression	SVM
Train Error	0.333 -> 0.358	0.294 -> 0.282	0.415 -> 0.419	0.316 -> 0.322	0.215 -> 0.234
Test Error	0.345 -> 0.371	0.301 -> 0.279	0.442 -> 0.442	0.298 -> 0.305	0.192 -> 0.213

Improvement for trees

No Improvement

Worse

Recursive Feature Elimination

- What is RFE?
- Performance with stratified k-fold:

	Decision Trees	Random Forests	K-nearest neighbours	Logistic Regression	SVM
Train Error	0.333 -> 0.289	0.294 -> 0.219	0.415 -> 0.321	0.316 -> 0.318	0.215 -> 0.215
Test Error	0.345 -> 0.263	0.301 -> 0.203	0.442 -> 0.313	0.298 -> 0.322	0.192 -> 0.204

Improvement
for trees

Minor/No
Improvement

No
Improvement

ANOVA F Measure

- What is ANOVA F Measure?
- Performance with stratified k-fold:

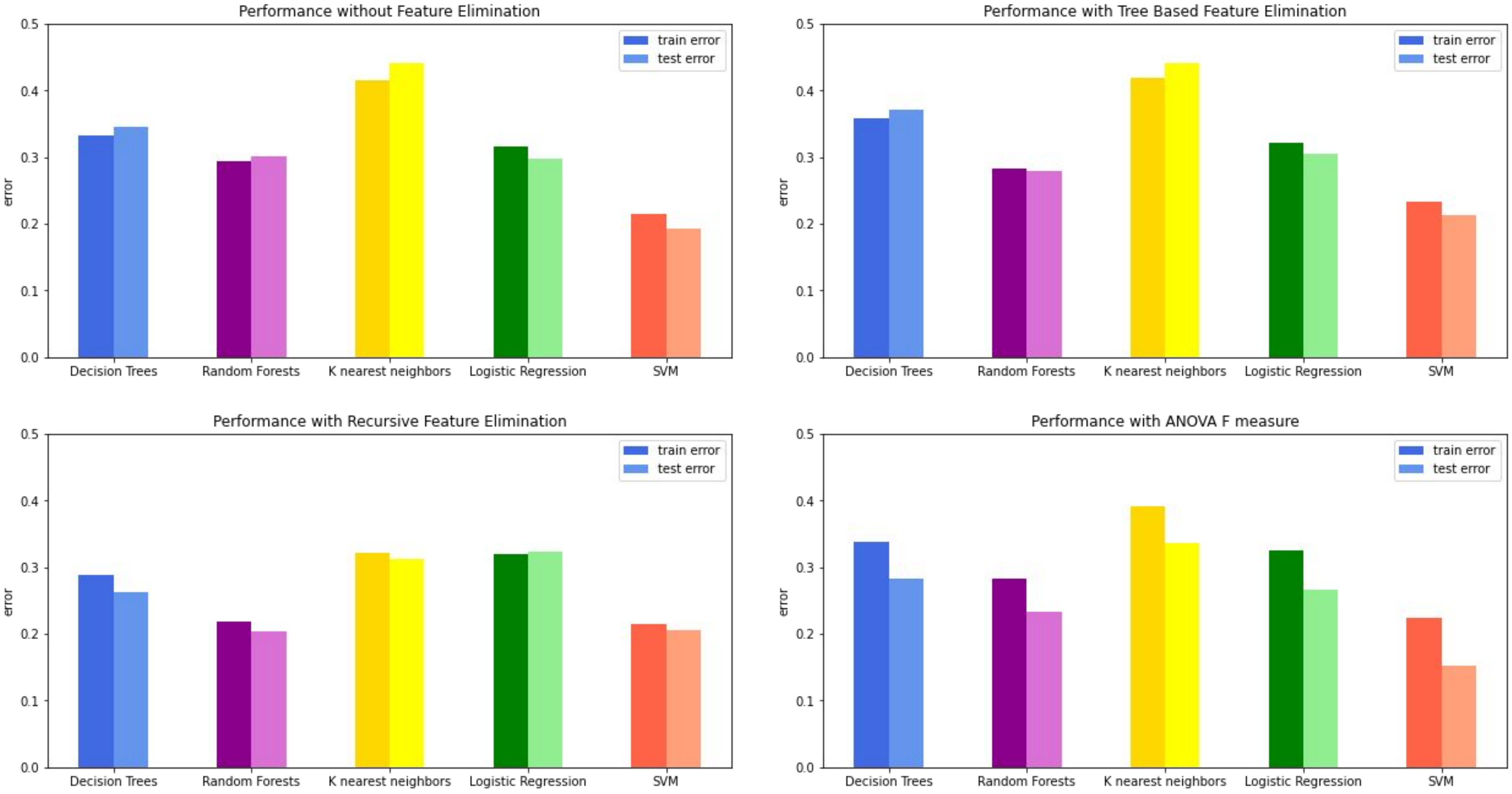
	Decision Trees	Random Forests	K-nearest neighbours	Logistic Regression	SVM
Train Error	0.333 -> 0.338	0.294 -> 0.283	0.415 -> 0.391	0.316 -> 0.324	0.215 -> 0.224
Test Error	0.345 -> 0.282	0.301 -> 0.234	0.442 -> 0.335	0.298 -> 0.266	0.192 -> 0.152

Minor/No Improvement

Minor Improvement

Minor Improvement

Performance of all Feature Selection



Performance without Feature Elimination

Performance with Tree Based Feature Elimination

Performance with Recursive Feature Elimination

Performance with ANOVA F measure

Experiment #2

Continued analysis of previously selected model

Sub-tasks:

- Additional data preprocessing
- Feature selection assessment
- Hyper-parameter tuning

Additional Preprocessing: Feature Matrix (X)

- Dealing with multiple artists on a single song
  - Dataset grouped multiple artists into a single string (“Artist1 - Artist2 - Artist3”)
  - We decided to split the artists into a list and encode whether they were present on any song
  - Eliminated “grouped” artists, resulting in fewer columns than previous experiment

- Retained all features except song title
  - Required dropping all rows with missing values
    - Resulted in fewer rows corresponding to certain genres
    - Next slide shows further genre grouping

Example:

ArtistA, ArtistB, & ArtistC made Song1 together.

Dataset previously considered “ArtistA - ArtistB - ArtistC” to be a single artist.

Splitting and encoding ensured that Song1 was attributed to each artist individually.

Additional Preprocessing: Target Labels (y)

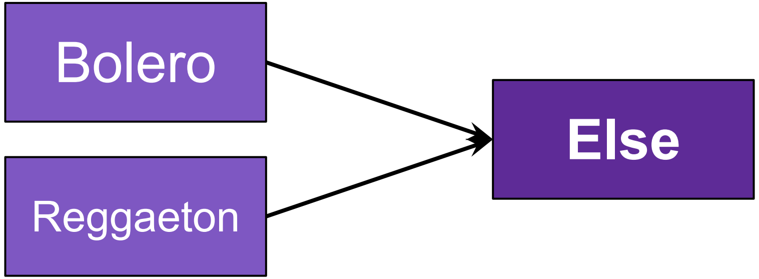
- Dropping non-English songs resulted in fewer entries for certain genres
  - Table to the right shows status of genres after dropping

- Decided to group some more genres into larger groups

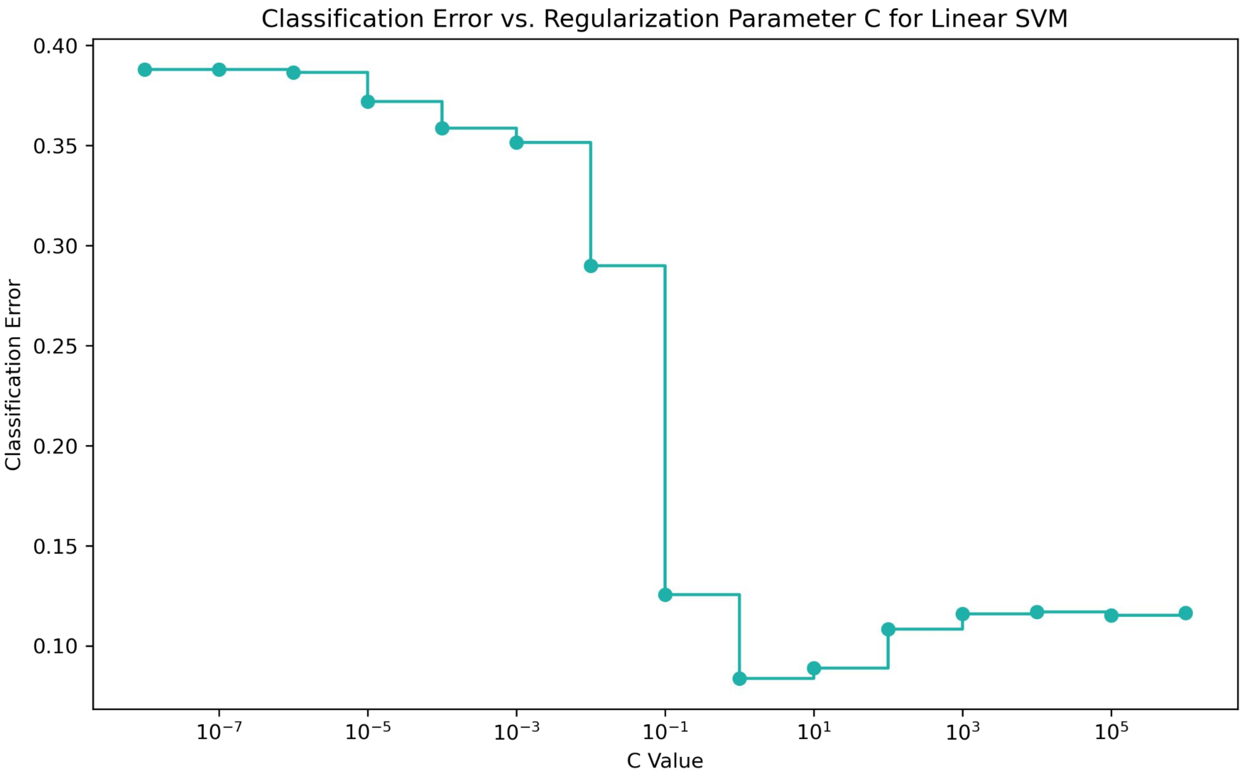
Genre	# of songs
hip hop	1241
pop	1155
dance/electronic	189
rock	177
r&b/soul	125
boy band	63
k-pop	42
else	36
indie	35
country	19
latin	9
funk	8
reggaeton	4
bolero	1

# of songs: 1

# of songs: 4



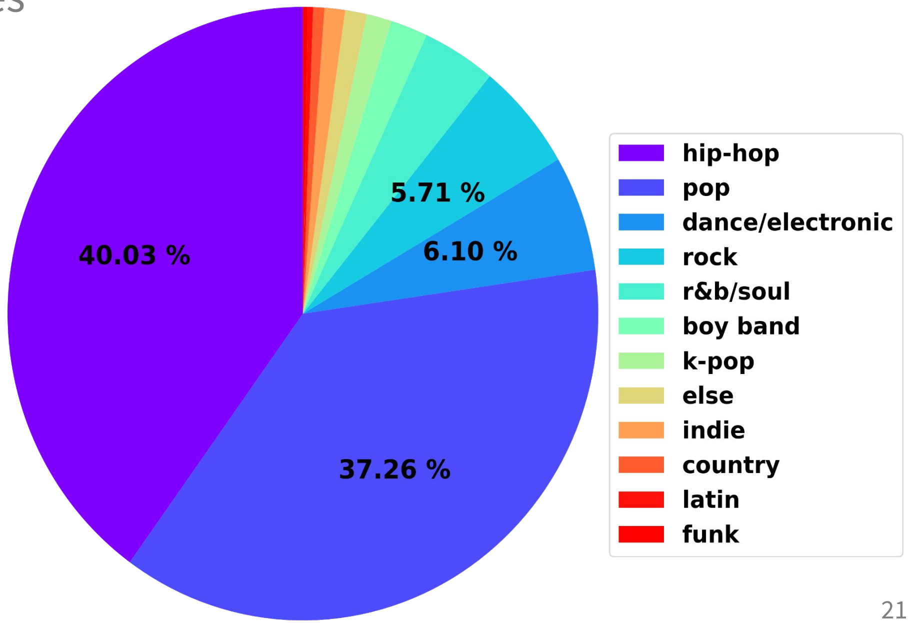
SVM - Regularization Parameter (C) Tuning



Conclusion

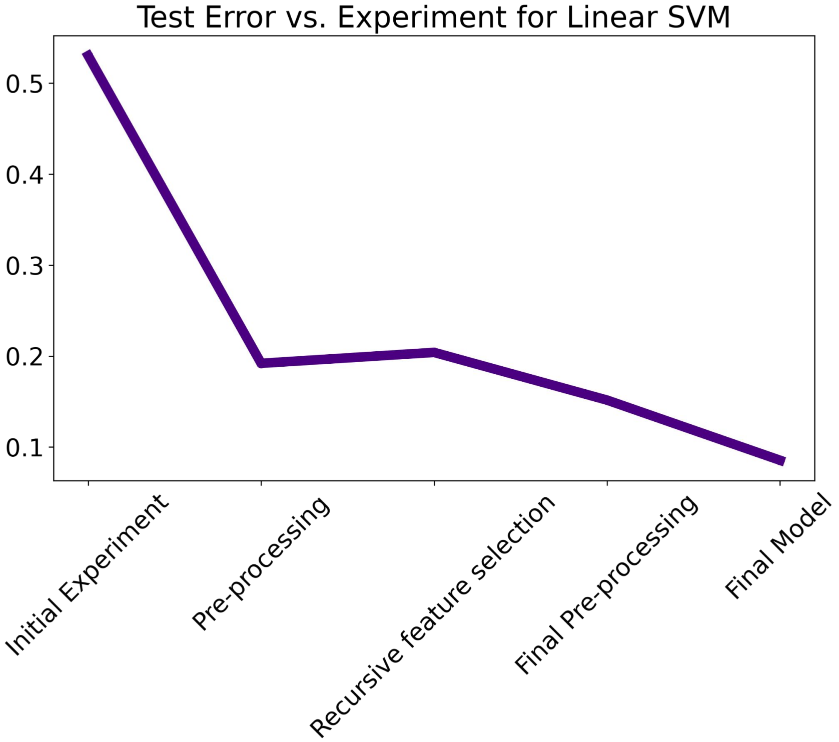
Breakdown of Final Dataset

- Dropped all rows with missing values
- Focused on songs from English-speaking countries
- One-hot encoding for non-numeric features
- Standard scaling for numerical features
- Re-labeled sub-genres into larger genres
- Utilized binary encoding for artists



Breakdown of Final Model

- Recursive Feature Elimination
- Stratified 5-fold cross validation
- Linear SVM Classifier
  - one vs. rest
  - L2 penalty
  - Hyper-tuned C of 1.0



0.0755

Final Model Test Error

Better than the initial project goal of 0.15 !

Reflection & Challenges

- Understanding and working with the dataset
- Determining how to pass the data to various Machine Learning models in such a way that they produce (relatively) well results
- Computational cost of various feature selection algorithms
- Version control and coordination between multiple coders

Additional Reflections

- Successful multi-class classification requires more examples
  - Can a genre in the test set be classified correctly if it was not present in the training set?
  - Side Experiment: A binary classification of Hip Hop (1155 examples) vs. Pop (769 examples) results in a consistent classification within 2% test error (with Linear SVM)

- Therefore, we **can** classify a song’s genre using its metadata

- Future Steps:
  - Gathering more examples and expanding beyond English-language songs
  - Exploring classification of genres using songs’ waveforms

Q & A

Thank you for listening!