Hit Song Science:

Modeling Weekly billboard 100 Chart Ranking with twitter Data

SI 699 Midterm Report | Team 6
Hsin-Yuan Wu | John Lee | Melody Chang

Background

- **Importance for Marketers/Business:**
  - Crowd wisdom gives estimates that are particularly beneficial in detecting trends
  - Improve product outreach by understanding different demographic groups
- **Importance for You:**
  - What's there better to keep things stimulated during COVID other than Music?
  - Hey, maybe you will become an artist one day!
- **Research Question: What are some important features that help predict weekly rankings of the top 100 songs on Billboard?**
  - Can we detect the rank movement of songs on the Billboard 100 chart using information from twitter (e.g., hashtag mentions, retweets)?
  - Is there a connection between how a song is performing today and how it performed in previous weeks (e.g., ranking patterns)?

- On January 6th, 2020, Mariah Carey's holiday hit "All I Want for Christmas is You" fell from Billboard's #1 spot to completely off the top 100 chart.

- What caused this flash drop?

Previous Work

- *Using Twitter to Predict Chart Position for Songs* (2020) by Tsiara and Tjortjis
  - Aimed to use Twitter data to predict the top 10 chart positions for songs on the Billboard Hot 100 charts.
  - Collected more than 1M tweets during Oct. and Nov. in 2018 via Twitter Search API.
  - Example techniques:
    - VADER's sentiment analysis for consolidating twitter text parameters (e.g. positive/negative/neutral tweets)
    - Ensemble Trees (e.g, Random Forest) and regression analysis (e.g., Support Vector Regression) for classifying ranks of the top 10 Billboard songs
- Suggest that there is a moderate correlation between the number of mentions of a song and its chart performance.
- No significant relationship was observed between the attention an artist gets on Twitter, even via emotionally charged tweets, and their tracks' success.

Challenges

- A lot of variation in features of a single song that make it difficult to extract metrics that represent an entire song
  - sentiments- positive, negative, neutral, etc
  - acoustic features
- Takes a tremendous amount of time (~30mins) to load in millions of twitter records from 1 day, let alone 3 months (13 weeks)

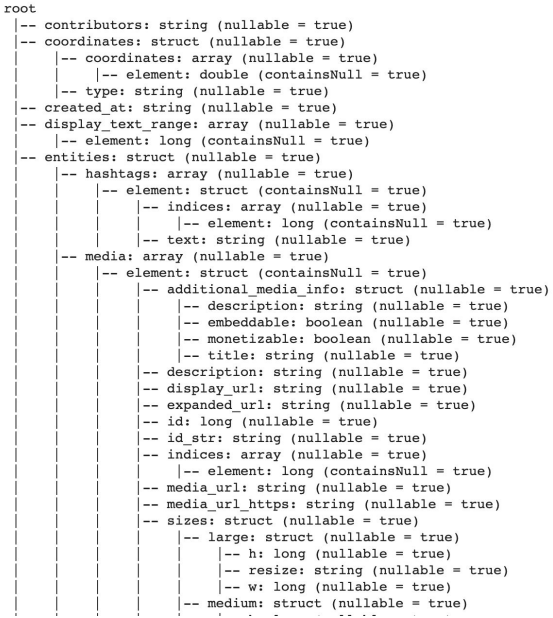
Data - Billboard Dataset

- A subset of weekly Billboard Hot 100 singles charts from data.world provided by Sean Miller.
- September 2020 to November 2020.
  - 3 months (13 weeks) of data - 1,300 records in total.
- Each row of data represents a song and the corresponding position on that week's chart.

id	date	rank	artist	song	rank_last_week	peak_rank	weeks_on_chart
3901	10/3/20	1	Cardi B Featuring Megan Thee Stallion	WAP	1	1	6
3902	10/3/20	2	BTS	Dynamite	2	1	4
3903	10/3/20	3	Drake Featuring Lil Durk	Laugh Now Cry Later	3	2	5

Data - Twitter Dataset

- Twitter Decahose Stream via MIDAS.
- Obtained data from August 29, 2020 to November 27, 2020.
- Each row in the dataset represents a tweet.
- More than 150 attributes available.
  - 38 main columns - majority of them are a nested array or a nested struct.
- For the purpose of this project, we pre-selected 28 columns to include in further analysis.



Partial schema of the Twitter Dataset

Methods

Feature Extraction

Billboard Dataset
Weekly chart rankings and song information.

+

Twitter Dataset
Extract features related to popularity, network patterns, textual components, etc.

Model Building

Feature Selection
Tree-based models to narrow down features.

Multi-class Classification
Support Vector Machine, Logistic Regression, Gradient Boosting, and Multi-layer Perceptron.

Model Optimization

Potential Methods
Hyperparameters tuning, data imputation, normalization, etc.

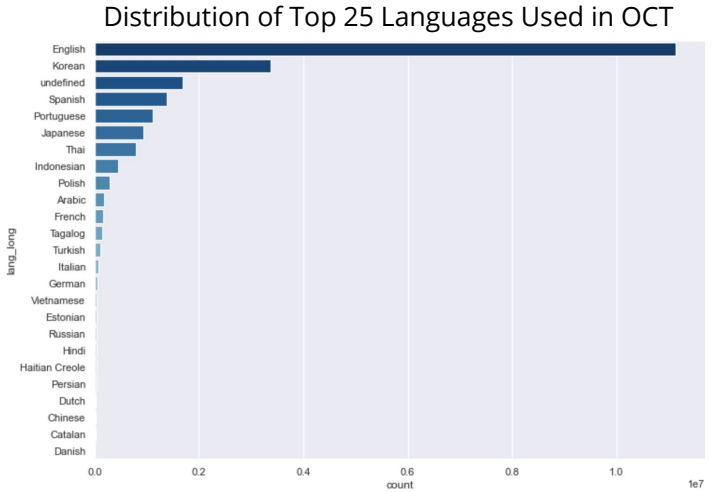
Evaluation
MAE, Confusion matrix, and AUC-ROC curve. Compared to baselines.

Data Preprocessing - Billboard

- Data type transformation and factorization.
- Transformed to a multi-class classification problem.
  - Tier 1: rank 1 to rank 20
  - Tier 2: rank 21 to rank 40
  - Tier 3: rank 41 to rank 60
  - Tier 4: rank 61 to rank 80
  - Tier 5: rank 81 to rank 100
- Obtained a list of unique song names and artist names in the three month interval to support preprocessing of the twitter data.

Data Preprocessing & EDA - Twitter

- Explored data in October
  - Approximately 1.3 billion records.
- Filtered for relevant records/tweets
  - Only include records where the tweet content contains song name or artist name from the Billboard list.
  - Ended up with 22,350,893 of records for EDA.
- Distribution of languages used
  - English is the most used language.
  - Many tweets with “undefined” language.



Data Preprocessing & EDA - Twitter

Explored Data By Week

- Top 10 most popular songs based on tweet counts
  - *Dynamite* by BTS dominated the ranking throughout October.

weeks	song_name	weeks	song_name	weeks	song_name	weeks	song_name	count
0.0	Dynamite	1.0	Dynamite	2.0	Dynamite	3.0	Dynamite	84468
0.0	Savage	1.0	ily	2.0	ily	3.0	ily	26807
0.0	ily	1.0	24	2.0	24	3.0	24	19556
0.0	All In	1.0	21	2.0	Lonely	3.0	Positions	18505
0.0	Lovesick Girls	1.0	Savage	2.0	Lovesick Girls	3.0	Golden	17645
0.0	24	1.0	Better	2.0	21	3.0	21	17433
0.0	21	1.0	Lovesick Girls	2.0	Savage	3.0	Better	10741
0.0	Ice Cream	1.0	Shut Up	2.0	Dreams	3.0	Cardigan	10354
0.0	Franchise	1.0	Wonder	2.0	Better	3.0	Done	6683
0.0	Like That	1.0	Done	2.0	Done	3.0	Wonder	5449

Explored Popularity Features

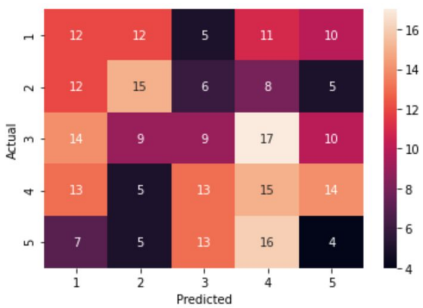
- Total number of retweet / reply / favorite / quote counts
  - Sparse features

song_name	sum(retweet_count)	sum(reply_count)	sum(favorite_count)	sum(quote_count)
Why We Drink	0	0	0	0
Why Would I Stop?	0	0	0	0
Throat Baby (Go B...	0	0	0	0
More Than My Home...	0	0	0	0
Hate The Other Side	0	0	0	0
Beers And Sunshine	0	0	0	0
Mr. Right Now	0	0	0	0
Steppin On N*ggas	0	0	0	0
One Of Them Girls	0	0	0	0
Whats Poppin	0	0	0	0
RIP Luv	0	0	0	0
Pretty Heart	0	0	0	0
Big, Big Plans	0	0	0	0
Party Girl	0	0	0	0
Brand New Draco	0	0	0	0
I Called Mama	0	0	0	0
I Love My Country	0	0	0	0
La Toxica	0	0	0	0
Take You Dancing	0	0	0	0
Money Over Fallouts	0	0	0	0

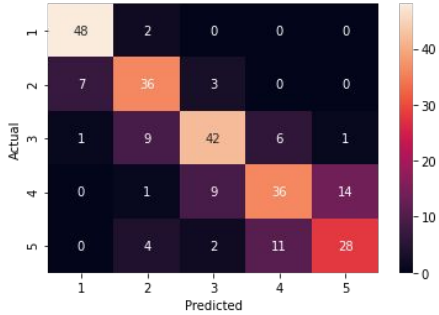
Baselines

- Baseline 1 - Dummy Classifier with ‘uniform’ strategy
  - The classifier generated predictions uniformly at random.
  - The random state of the classifier was set to 0 to ensure the results are reproducible.
- Baseline 2 - SVC
  - We trained a SVC classifier with the default parameters: ‘rbf’ (radial basis function kernel) as the kernel and ‘ovr’ (one-vs-rest strategy)
- Evaluation: MAE, Precision, Recall, F1, Confusion Matrices

Evaluation	**MAE**	**Macro Precision**	**Macro Recall**	**Macro F1 Score**
Baseline 1: **Dummy**	1.46	0.21	0.21	0.21
Baseline 2: **SVC**	0.319	0.73	0.74	0.73



Dummy

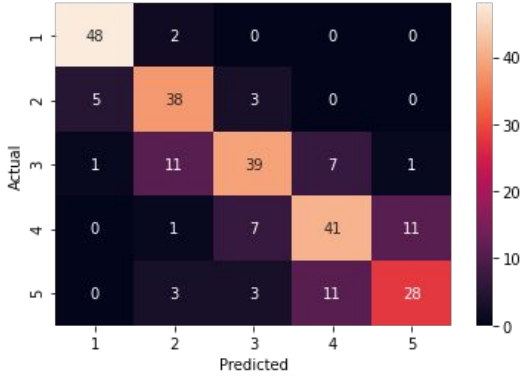


SVC

Additional Results

- Multilayer Perceptron (MLP) classifier
  - default parameters with activation function of ReLu
  - Compare to baseline: a little better than SVC

Evaluation	**MAE**	**Macro Precision**	**Macro Recall**	**Macro F1 Score**
MLP	0.3	0.74	0.75	0.75
SVC	0.319	0.73	0.74	0.73

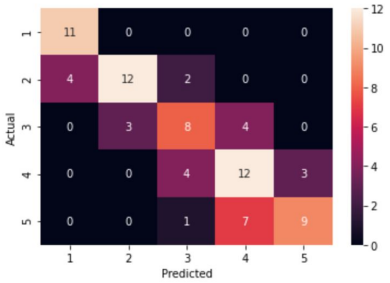


MLP Confusion Matrix

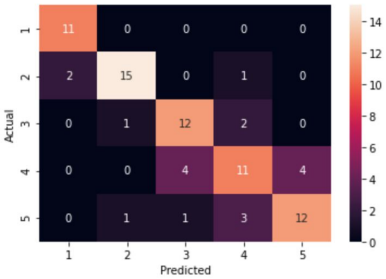
Additional Results - Features Added

- We extracted two features from the Twitter dataset and added to the Billboard dataset for model building in the subset of the dataset
  - Period: October 2020
  - count_tweet_song: Number of tweets relevant to the song in the time interval
  - count_tweet_artist: Number of tweets relevant to the artist in the time interval
- SVC and MLP
  - With small subset of data, we have a better performance on MLP (F1: 0.76 vs 0.77), but worse on SVC (F1: 0.73 vs 0.66).

Evaluation	**MAE**	**Macro Precision**	**Macro Recall**	**Macro F1 Score**
SVC	0.36	0.67	0.67	0.66
MLP	0.29	0.77	0.78	0.77



SVC



MLP

Conclusions

- Feature extraction from the Twitter dataset is possible.
- Our proposed approach is feasible.

Next Steps

- Feature Extraction
  - Consider whether urls or emojis in tweets could affect viewers' perception regarding the songs
  - Retrieve tweet contents features such as the average length of the tweets for each song/artist and conduct TF-IDF analysis on the text
- Machine Learning Technique
  - Utilize tree-based models, like Random Forest and Extremely Randomized Trees, to help us narrow down and select relevant features
- Model Optimization
  - Explore regularization and/or normalization, detection and removal of outliers, imputation of missing values, feature selection, and cross-validation to avoid overfitting or underfitting the models
  - Other potential neural network models

Thank You