Spotify Tracks Popularity Analysis

Jongwan Kim | Richard Lin | Lingfei Zhao

Agenda

I	Problem & Motivation
II	Dataset Structure
III	Data Preprocessing
IV	Methodology
V	Results
VI	Conclusion & Next Steps

Problem & Motivation

What features make a song more popular on Spotify in general and within its own genre?
How well can we predict a song's popularity?
How can we enhance Spotify's recommendation system?

Why are these questions important?

1. Spotify has **356 million users, 70 million tracks, and 1.2 million artists**
2. The ability to **forecast a song's success** is crucial for stakeholders in the music industry
    - Marketing strategies
    - Resource allocation
    - Content curation

Numerous potential applications and benefits

1. Provides both artist and record labels **actional insights**
2. Empowers stakeholders to make **data-driven choices** for markets
3. Can **improve Spotify's playlist recommendations** and increase user satisfaction

Dataset Structure

Data Source

Dataset sourced from Kaggle: contains almost **90,000 unique tracks** on Spotify and covers **125 different genres**.

Structure

- 84,316 unique track ID's, each with corresponding features such as artist, genre, duration, danceability, and loudness
    - 8 nominal variables, 1 ordinal variable, 1 discrete variable, and 9 continuous variables
- **Popularity Score *(target prediction variable):*** calculated by an outside algorithm based on the **total number of plays** the track has had and how **recent** these plays were

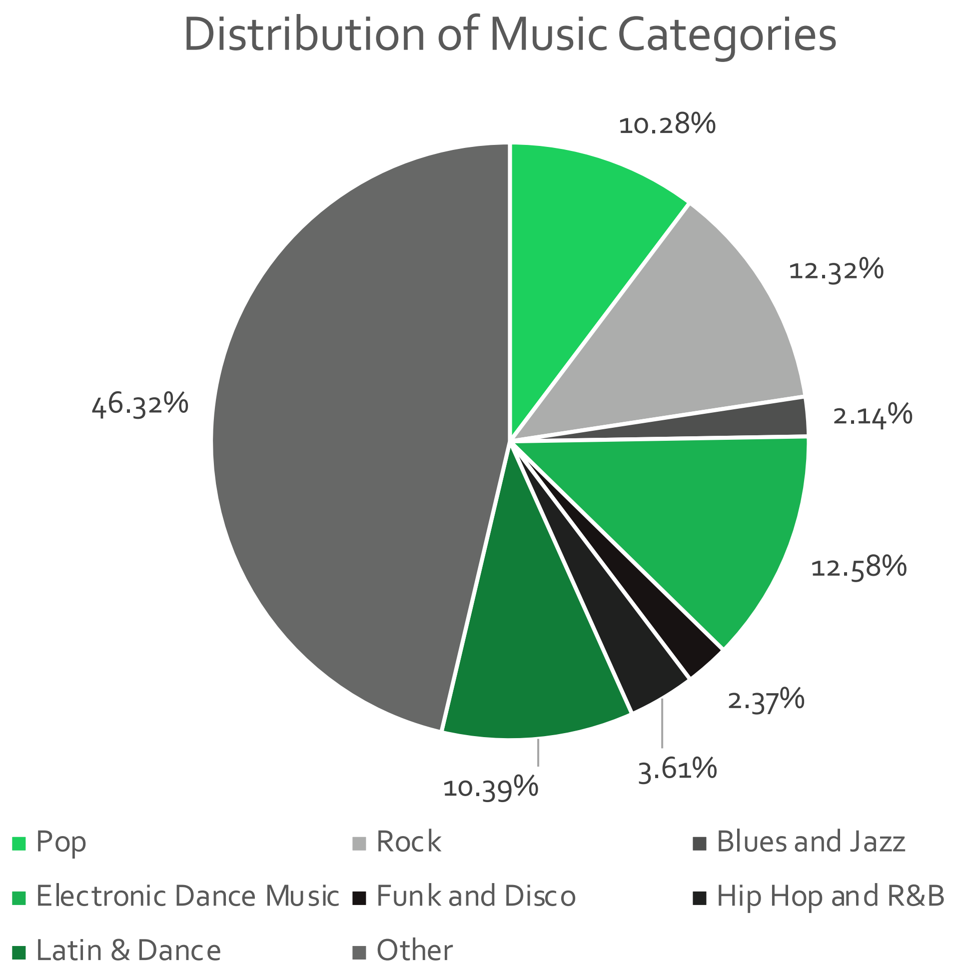
Justification

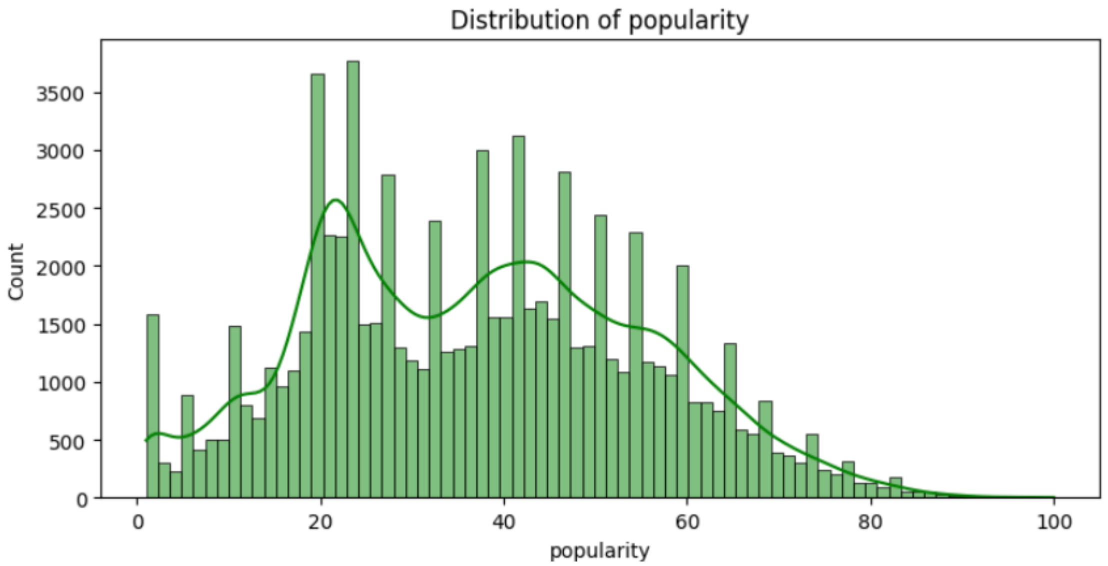
We can develop a robust predictive model with clear and easily understandable results due to the following:

- Large number of entries in the data source
- Comprehensive and well-documented feature set
- Spotify is one of the most popular and widely used platforms, so it is representative of the general population

Data Preprocessing

- Removed null and duplicate values
- Removed popularity scores of o
- Analyzed Song Descriptors: energy, loudness, danceability, tempo,
- Changed specific track genres into broader music categories: Electronic Dance Music, Hip-Hop, Rap, Latin and Dance, Rock

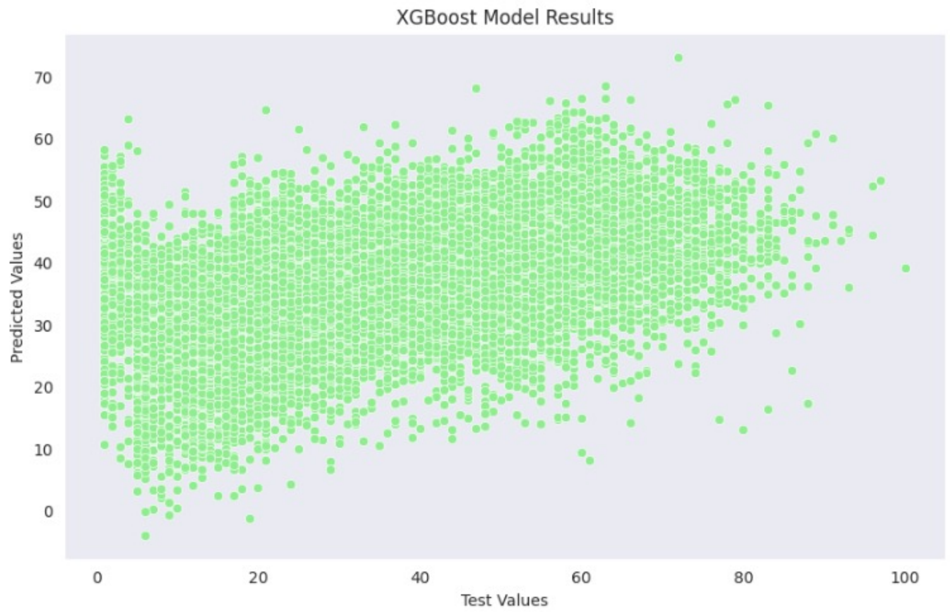


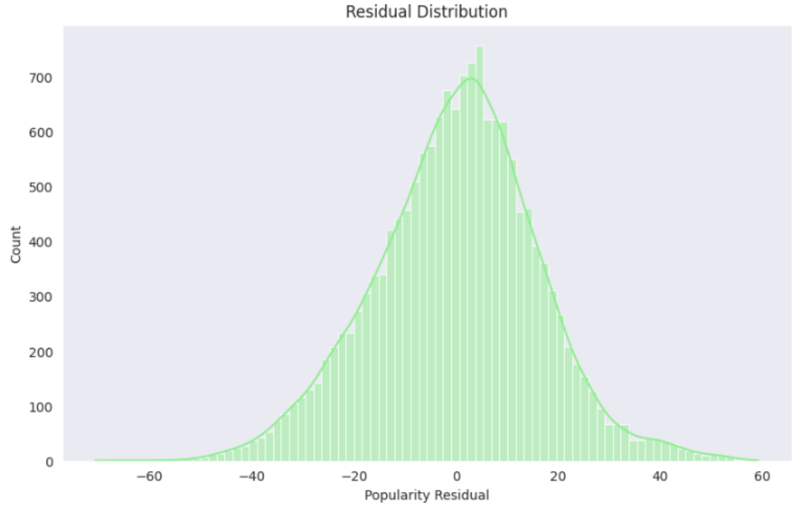


Methodology – Regression

- Find impact of song features on popularity score
- Train multiple Regressive Models to find optimal R-squared Score
- XGBoosted Decision Tree highest R-Squared Score in comparison

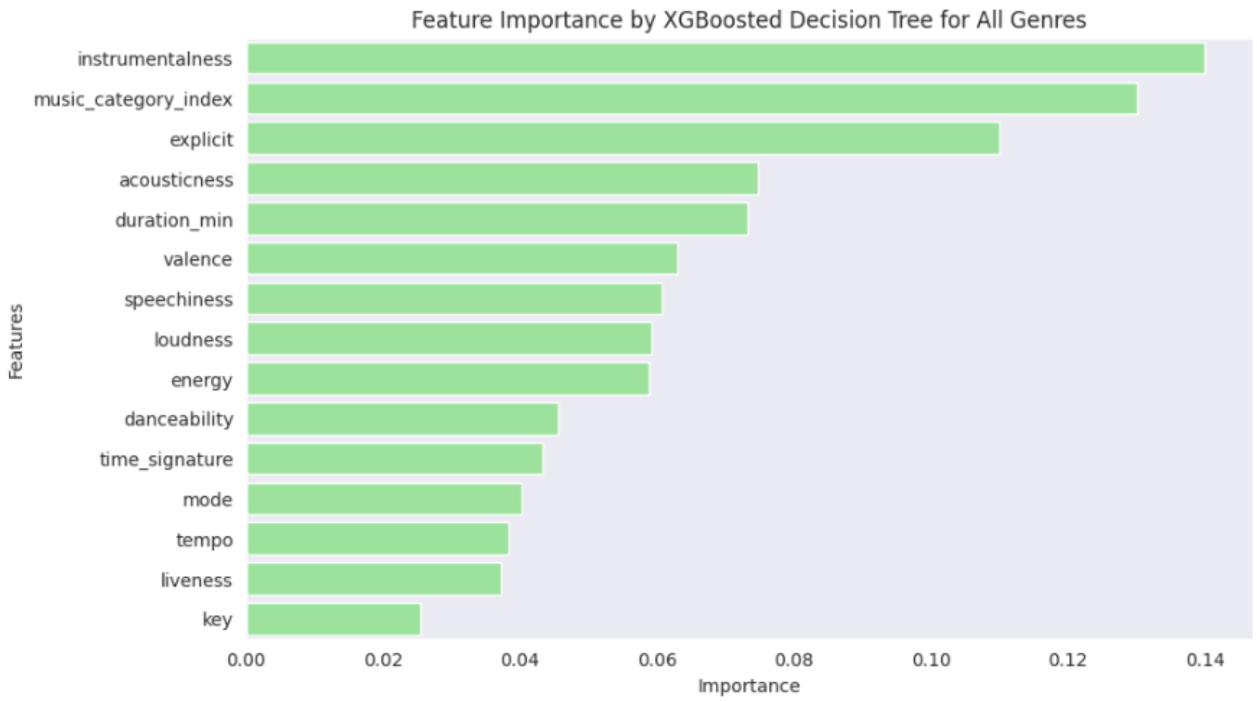
Model	MSE	R-Squared
Linear Regression	306.29426	0.08184
Ridge Regression	306.29431	0.08184
Lasso	309.48846	0.07227
Decision Tree Regressor	488.67806	-0.46488
XGBoost Regressor	**257.03304**	**0.22951**
Polynomial 2 degrees	284.70259	0.14656
Polynomial 3 degrees	277.22900	0.16897

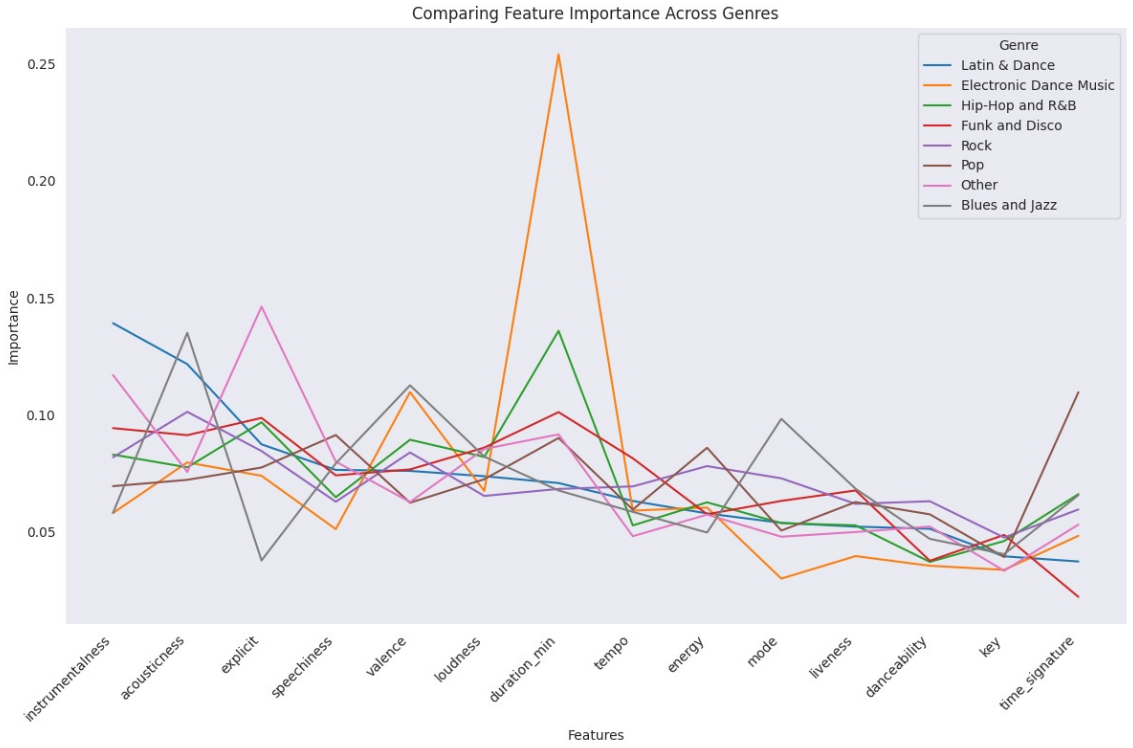




Methodology – Feature Importance

- Gradient boosted decision tree to calculate the importance of a feature across all songs
- Split training data by music categories to see observe changes in important music descriptors across genres
- Top feature for each music category was different based on the characteristics of the genre

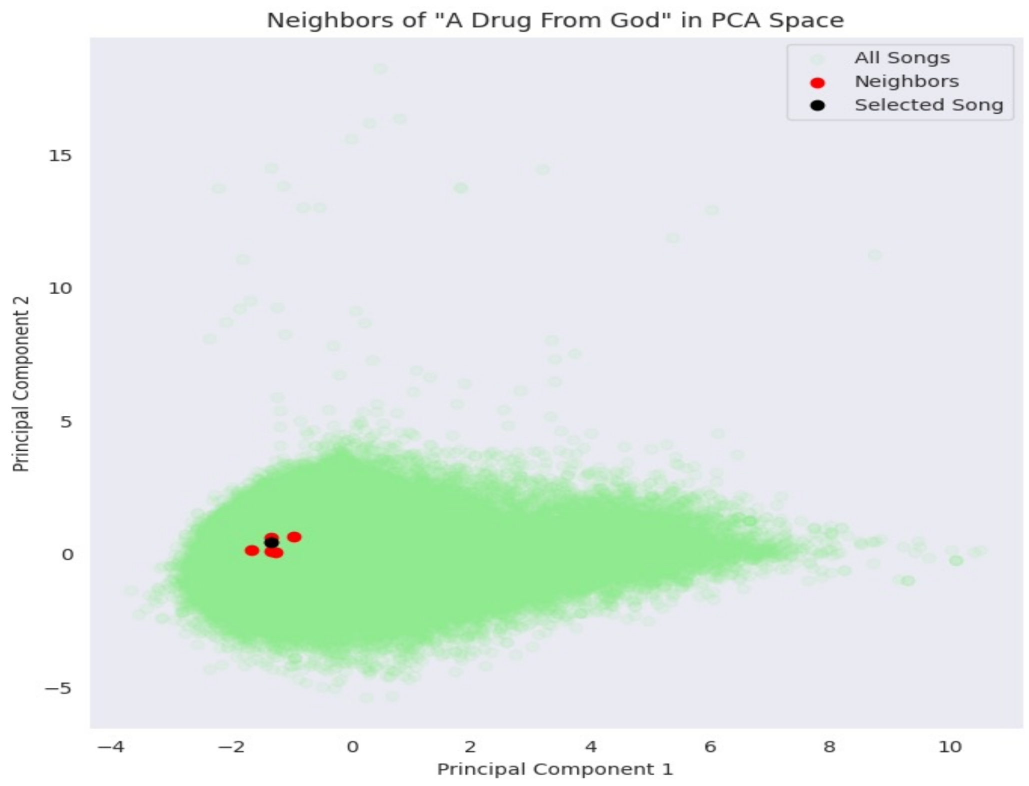


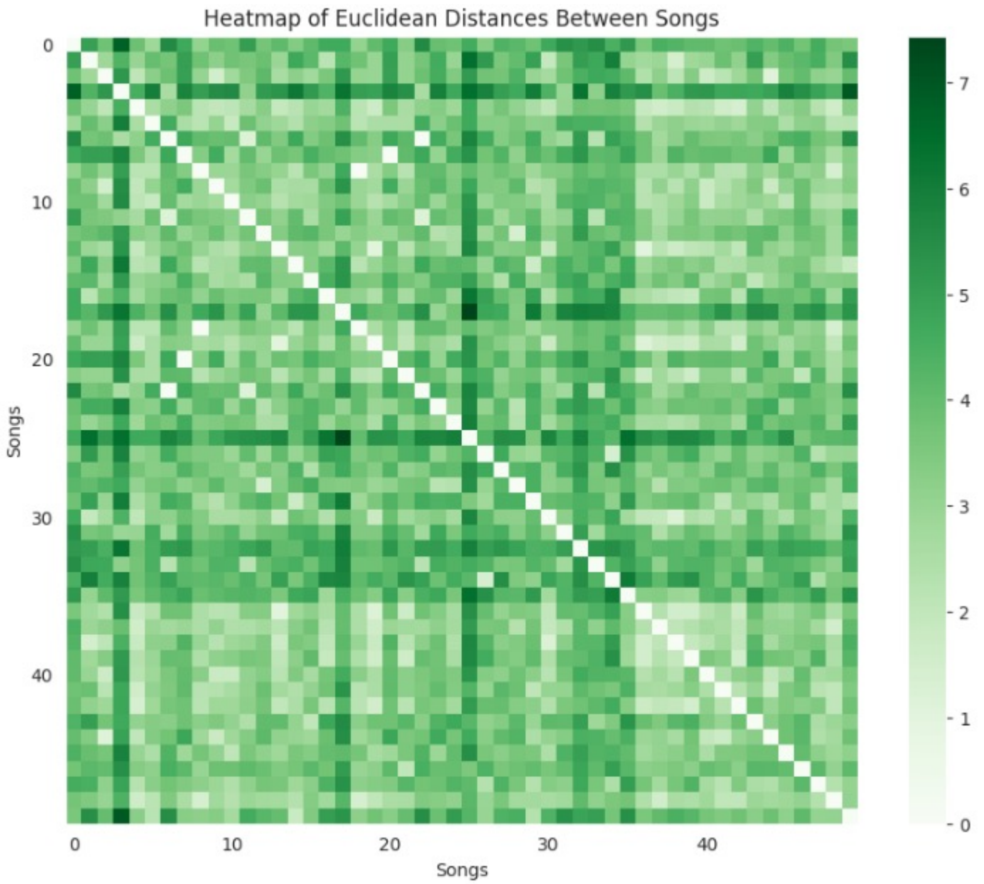


Question | Data Source | Data Preprocessing | Methodology | Results | Conclusion

Methodology – KNN and Principal Component Analysis

- Incorporated K-nearest neighbors algorithm and principal component analysis to make song recommendations based on user input
- Generated a heatmap of pairwise Euclidean distance between the first 50 songs in the dataset: Song #17 and Song #25 are the most different

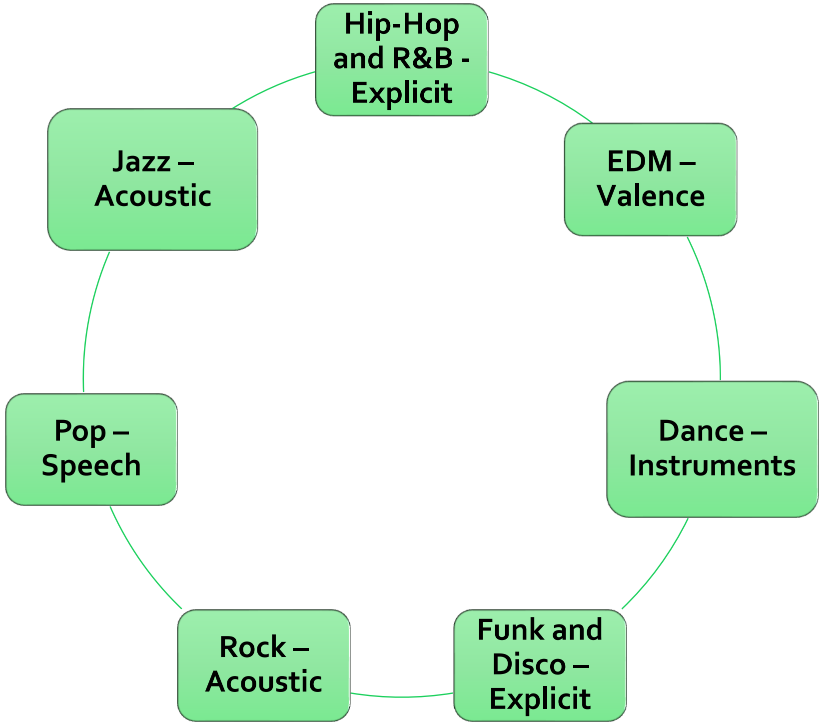


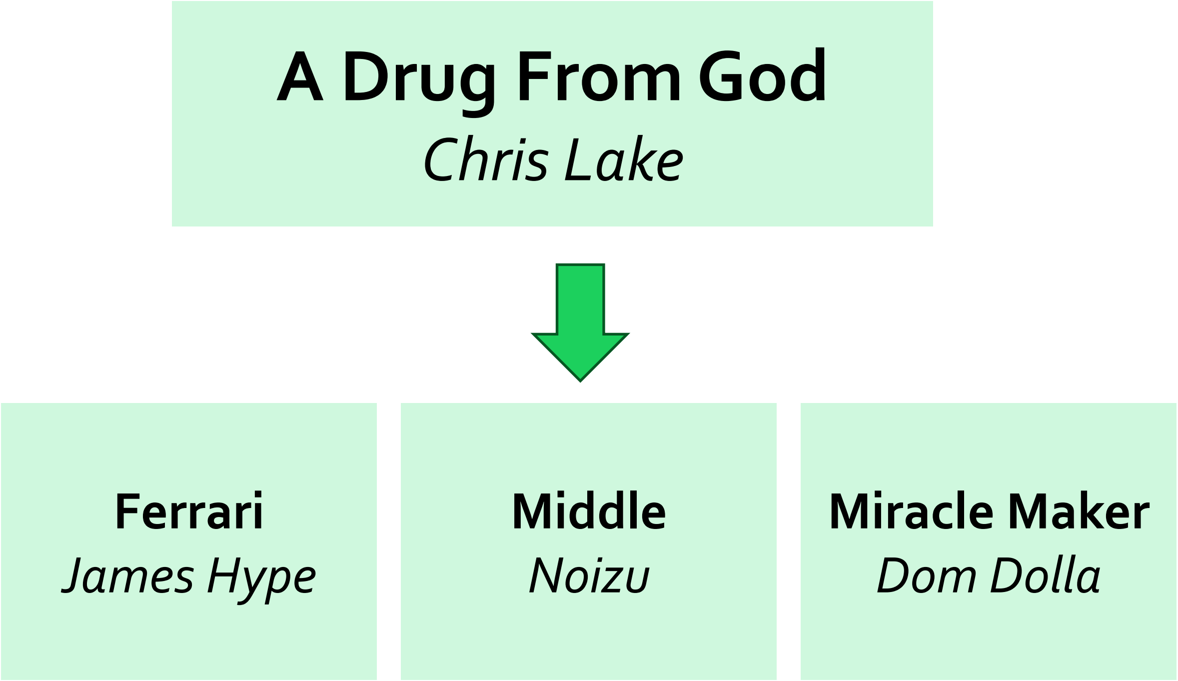


Question | Data Source | Data Preprocessing | Methodology | Results | Conclusion

Results

- Feature Importance by categories DO correlate to the characteristics of that music
- Use XGBoosted Decision Tree to gauge relative popularity of a song
- Generated tools to help artists produce hit songs more effectively, and provide Spotify with ways to increase user engagement via a song recommendation system





Conclusion & Discussion

Next Steps for Future Iterations

- Instead of dropping *artist, track_name, and album_name*, use NLP to extract information
- Analyze change in popularity of genres or songs over time
- Improve recommendation schema
  - Currently incorporates content-based filtering, which can cause "over-specialization". Mitigate by using genetic algorithm, which allows for more diverse recommendations

Fairness/Weapon of Math Destruction

Our model has the potential to become a **weapon of math destruction**

- Outcomes are not hard to measure, but can be subjective based on how popularity is calculated
- Predictions affect results and can create negative feedback cycles

Future iterations must consider fairness and the existence of demographic biases

Thank you!

References

MaharshiPandya. (2022). 🎹 Spotify Tracks Dataset [dataset]. https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset