Spotify Final Project

Yan He, Peter Valentine, Amany Yaakoub, Leona Yang

Agenda

- Selected Topic and Reason
- Description of Source Data
- Questions to be Answered
- Data Exploration Phase
- Analysis phase
- Dashboard Plan
- Recommendation

Selected Topic and Reason

- Dataset of Spotify tracks Predict the popularity of a track
- High interest in music and to better understand how a track becomes a hit

Description of Source Data

Popularity	Duration	Key	Mode	Time Signature
Acoustic	Energy	Liveness	Speech	Valence
Danceability	Instrumental	Loudness	Tempo	

Questions to be Answered

- What features are the most predictive in determining track popularity?
- What are the optimal combination of features for popular tracks?
- Are there any differences across genres?

Data Exploration Phase

- Used Python (Pandas) to examine records, for example:
  - How many tracks were in each type of genre?
  - Which ones were underrepresented?
  - Which features likely had no bearing on whether it was a hit (such as key or tempo)?
  - How many tracks could be removed from various genres that were not likely to be musical “hits” such as comedy?

Database ERD

Tools and Languages

- HTML
- JS
- CSS
- Flask
- SQLAlchemy
- Tableau
- Python
- Pandas

Analysis Phase

- Started with ETL
- Random Forest Machine Learning Model (used to help identify predictive features)
- Data will be analyzed/visualized using Tableau
- Website to be developed with dashboard and a prediction tool to predict tracks to be a hit or not?

Machine Learning Phase

Target

To build a machine learning model to interpret a track is a hit song or not. After inspect the raw data, we decided to determine a track is a hit song or not base on it's popularity, if it's popularity is over 50 than it will be considered as a hit song.

Module Selection

- Neutral Network
- Logistic Regression
- Random Forest

Results of Analysis

	Accuracy Score
Neutral Network	75%
Logistics Regression	74%
Random Forest	82%

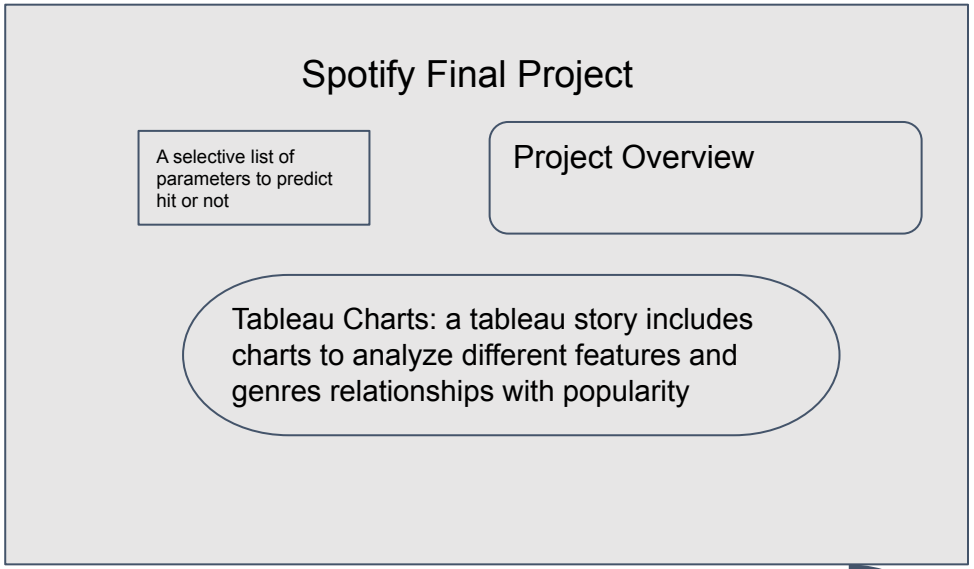
Results of Analysis

- According to the accuracy scores, we can see that the Random Forest model has the highest accuracy of 82%. We can say that the Random Forest model is pretty accurate to determine whether a track is a hit song or not.
- From the importance test, we can see that 'acousticness', 'loudness', 'danceability', 'valence', 'tempo', 'speechiness', 'energy', 'liveness', 'instrumentalness' are of similar importance, being 0.10 or more, and 'explicit' is not as important as other features with only 0.007.
- The importance between different genres are ranging from 0.0027 to 0.0141, with genre_Jazz_Blues being the lowest and genre_Pop being the highest

Dashboard Plan

- Several dashboards to be done in Tableau
  - Artists dashboard bar graph
  - Features vs popularity scatter graphs and line chart
  - Average popularity vs genre bar graph
  - Top 100 artists (using sum)

Sample Dashboard



Recommendation for Future Analysis

- Larger Datasets
- Different years
- More ML models

Thank you