Data Science Capstone | Project 2

Mark Hassan

December 9, 2025

Business Scenario

Overview

In my scenario, the customer of this project is FutureProduct Advisors, a consultancy that helps their customers develop innovative and new consumer products. FutureProduct's customers are increasingly seeking help from their consultants in go-to-market activities.

FutureProduct's consultants can support these go-to-market activities, but the business does not have all the infrastructure needed to support it. Their biggest ask is for a tool to help them find interesting, up-and-coming music to accompany social posts and online ads for go-to-market promotions.

Stakeholders

- FutureProduct Managing Director: oversees their consulting practice and is sponsoring this project.
- FutureProduct Senior Consultants: the actual users of the prospective tool. A small subset of the consultants will pilot the prototype tool.
- My consulting leadership: sponsors of this effort; will provide oversight and technical input of the project as needed.

Primary Goals

1. Build a data tool that can evaluate any song in the Billboard Hot 100 list and make predictions about:
    a. The song's position on the Hot 100 list 4 weeks in the future
    b. The song's highest position on the list in the next 6 months
2. Create a rubric that lists the 3 most important factors for songs' placement on the Hot 100 list for each hear from 2000 to 2021.

Dataset

Billboard Hot 100 weekly charts (Kaggle)

I've chosen this dataset because it has a direct measurement of song popularity (the Hot 100 list) and because its long history gives significant context to a song's positioning in a given week.

The features list gives a wide range of song attributes to explore and enables me to determine what features most significantly contribute to a song's popularity and how that changes over time.

This dataset has two parts:

HotStuff: the songs in each week's Billboard Hot 100 list from Aug 1958 through May 2021

- 11 columns x 328,000 rows
- Source: Billboard
- Key features:
  - Week ID
  - Song name
  - Performer name
  - Cumulative weeks on chart
  - Previous week position

Hot100AudioFeatures: characteristics of each song that has appeared on the Hot 100, such as performer name, genre, and characteristics such as danceability, energy, etc.

- 23 columns x 29,500 rows
- Sources: Spotify, Billboard
- Key features:
  - Song name
  - Performer name
  - Spotify_genre
  - Duration (ms)
  - Tempo
  - Key

A machine learning approach is well suited to this project because of the large size of the dataset and in particular because of the large number of features. Other modeling approaches would have more difficultly identifying which features are contributing to outcomes, and to what degree.

Success Measures

1. The prediction model achieves 90%+ accuracy on an unseen dataset.

2. The prediction model is efficient enough that it can return results on a given dataset in less than 5 minutes.
3. The prediction model is user friendly enough that a consultant at the customer company can learn how to use it with 30 minutes of training and/or guidance.

Problem Solving Process

1. Data acquisition and understanding
    a. The dataset will be simple to pull from Kaggle
    b. I'll explore which features are missing by time period—I anticipate using data from 2000 forward but will see if a different time period gives higher-quality data
    c. I plan on visualizing:
        i. Genres, tempos, lengths of songs that moved 5+ positions week to week, per year
        ii. Genres, tempos, lengths of songs that reached the top 25 positions, per year
2. Data prep and feature engineering
    a. I'll make sure to address missing values (impute, remove, etc.) for the features and time periods that the modeling will cover
    b. I'll begin by including a wider range of features, then narrow the features based on those that are most influential
3. Modeling approach
    a. I'll evaluate these algorithms: boosting ensembles, kNN, RNN
    b. I'll use k-fold to cross validate
    c. I'll tune hyperparameters using random search then grid search
    d. I'll use accuracy as my evaluation metric, since I don't expect the data to be significantly imbalanced, and there is no particular penalty for either false positives or false negatives
4. Interpreting and communicating results
    a. I'll prepare a PowerPoint presentation to provide an overview of the results
        i. This will include a table of the most influential factors in song popularity by year, and a summary of the best-performing prediction model
        ii. I plan on visualizing feature importance by showing how feature importance varies model to model, and as the number of features are narrowed.
        iii. I'll avoid technical terms to the greatest extent possible. I will include a brief overview of how each model works in simple terms, using analogies when possible.
    b. I'll prepare a Word/PDF overview of how the prediction model works for use by the customers' consultants.

Timeline and Scope

Timeline

	Start	End	Hours	Notes
Problem formulation + dataset finalization	12/8	12/9	6	
Exploratory data analysis	12/9	12/10	8	
Data preprocessing	12/10	12/10	6	
Model development	12/11	12/11	8	
Model evaluation/ refinement	12/12	12/12	6	
Documentation and reporting	12/12	12/13	4	
Final review and submission	12/13	12/13	6	

Scope

In Scope

- Song popularity tool, which makes 2 predictions for any song on the Billboard Hot 100:
  - Song ranking 4 weeks in the future
  - Highest ranking the song will achieve in the next 6 months
- A list of the most influential factors for song rankings, each year 2000 - 2021

Out of Scope

- Predictions for songs not in the Billboard Hot 100
- Factors or predictions for years before 2000