Spotify Song Popularity

Jessica Hamilton, James Matlock, Bailey Perosa, Chidi Ukaegbu

***Abstract* – This proposal details what our project is about, our methodology, and our expected outcomes. The goal of the project is to classify song popularity based on known song attributes.**

I. INTRODUCTION

The primary objective of this project is to classify song popularity based on song attributes such as danceability, acousticness, and loudness. Popularity in this project will be explained on a scale from 0-100 and defined by the total number of plays a song has had and how recent those plays are (songs with more current plays will be classified as more popular than songs that were played a lot in the past). The motivation for this project is curiosity surrounding what song attributes most contribute to popularity. This knowledge can be used by radio stations and DJs to better serve a general audience. Artists could also use this information to make creative decisions for the success of future works. From a business perspective, this information can also promote a positive consumer experience.

II. DATA

The data being used in this project comes from the Spotify Web API. The data contains almost 175,000 rows and 19 columns of song attributes (acousticness, artists, danceability, duration ms, energy, explicit, id, instrumentalness, key, liveness, loudness, mode, name, popularity, release date, speechiness, tempo, valence, and year). All columns except for artists, id, name, and release date are numeric. Of these numeric columns, most contain values between 0-1 while only six columns contain values greater or less than 1 (duration ms, key, loudness, popularity, tempo, and year). The artists column contains a list of all the artists featured on a song. The id column contains a unique identifier for each song in the dataset. The name and release date columns provide the full name of each song as well as the month, day, and year it was released.

We will be using the popularity column to classify on by creating groups based on the numeric value of popularity (e.g., 1-33 = not popular, 33-66 = popular, 66-100 = very popular). We will remove the id column since we do not want this included as an explanatory variable in our models. We may also discretize other numeric columns so our models will run more efficiently.

III. TEAM RESPONSIBILITIES

Jessica and Bailey will be responsible for importing the data, as well as coding, running, validating, and gathering insights for the naïve, decision tree, and random forest models.

Reagan and Henry will be responsible for coding, running, validating, and gathering insights for the boosted tree, regularized logistic regression, and k-nearest neighbors models.

All members will work on data cleaning, drawing overall conclusions by model comparison, and the final report and presentation.

IV. TIMELINE OF MILESTONES

Below shows a general timeline for the project:

Week 1: Import and clean the data. Split the data into train and holdout.

Week 2: Create models (start with naïve).

Week 3: Tune models.

Week 4: Gather insights from models (e.g., variable importances). Create summary table of models. Pick best model using 1 standard deviation rule.

Week 5: Draw conclusions on which factors are most important in determining the popularity of a song.

Week 6: Create visualizations of the most important factors.

Week 7: Construct final presentation and report.

V. EXPECTED OUTCOME

A. Expected outcome as a team:

- We expect to gain a deeper knowledge of different predictive modeling for classification based on varying factors.
- We all are familiar with model building in R, and we hope to become more familiar with the model building process in Python.

B. Expected outcome of the project:

- We are expecting to build a model to predict song popularity correctly over 50% of the time, for both the training and the holdout sets.
- We expect the model to generalize well to new data.
- We expect the model to have some interpretability for the purpose of providing a list of factors that influence song popularity the most.
- We expect every model to perform better than the naïve model.
- We expect to provide meaningful visualizations of the most important factors.
- We expect to be able to communicate results clearly and concise in laymen’s terms, both written and orally.