Predicting BillboardHot100 Chart Entry for Songs

---

1. Abstract

This project explores the potential to predict whether a song will enter the BillboardHot100 using audio features from Spotify.Leveraging a dataset of 41,106 songs, we examined features such as danceability, energy, and valence to determine their influence on chart success. We employed machine learning models, including Logistic Regression, Random Forest, and Support Vector Classifier (SVC), with Random Forest emerging as the best-performing model after hyperparameter tuning, achieving an accuracy of approximately 79%. This project provides insights into the characteristics associated with hit songs and demonstrates the feasibility of predictive models in the music industry.

---

2. Introduction

The music industry continuously strives to understand what makes a song popular. Predicting whether a song will chart on the Billboard Hot 100 can offer valuable insights for producers, artists, and marketers. This project aims to build a model that classifies songs as “charting” or “non-charting” based on audio features like danceability, energy, and valence. We used Logistic Regression, RandomForest, and SVC for classification, tuning each model to maximize accuracy. The input to the model consists of acoustic features provided by Spotify, while the output is a binary classification indicating Billboard chart entry status.

3. Dataset and Features

- **Dataset**: The dataset consists of **41,106 songs** from the 1960s to the 2010s, labeled as charting (1) or non-charting (0). Each song includes various audio and metadata features.
- **Dataset link**: [spotify-dataset-kaggle]
- **Features**:
  - **Danceability**: A measure of how suitable a track is for dancing.
  - **Energy**: Indicates the intensity and activity level of the song.
  - **Valence**: Reflects the positivity of the track.
  - **Loudness**: The overall volume of the song in decibels.
  - **Tempo**: The speed or beat frequency, in beats per minute.
- **Preprocessing**: We removed unnecessary columns, standardized numeric features, and filled missing values with appropriate statistics. Additionally, we created a `duration_minutes` feature by converting the `duration_ms` column.

4. Methods

We implemented three classification models:

- **LogisticRegression**: This baseline model provided insight into the data's separability.
- **Random Forest**: Known for high accuracy in binary classification tasks, RandomForest also offers feature importance, allowing us to interpret influential factors.
- **SupportVectorClassifier(SVC)**:SVC was chosen for its effectiveness in high-dimensional spaces.
- **Hyperparameter Tuning**: We applied Randomized Search for hyperparameter optimization on both Random Forest and SVC. Parameters optimized included `n_estimators`, `max_depth`, `C`, and `gamma`.

5. Experiments, Results and Discussion

- **HyperparameterTuning**:RandomizedSearchwith5-fold cross-validationprovidedthebestparametersforRandomForest (`n_estimators`=200, `max_depth`=30) and SVC.
- **EvaluationMetrics**:Accuracywasusedastheprimarymetric.A confusionmatrixwasalsousedtoevaluateprecisionandrecallfor each model.
- **Results**:
  - **LogisticRegression**: 73%accuracy.
  - **RandomForest**: 79%accuracy, making it the best-performing model.
  - **SVC**: 78%accuracy,with results close to RandomForest.
- **Discussion**:RandomForest provided the highest accuracy,likelydue to its ability to capture complex feature interactions. The SVC model's performanceindicatesthatnonlinearrelationshipsarerelevantforthis dataset. Cross-validation was used to prevent overfitting, but future work could involve additional regularization techniques.

6. Conclusion and Future Work

This project demonstrates the potential of machine learning to predict BillboardHot100entries,withRandomForestyieldingthehighestaccuracy. Danceability, energy, and valence were identified as the most significant predictors. In future work, we could expand the dataset to include more recenttracks,incorporatetextualanalysisofsonglyrics,orexploreneural networks to capture even more complex feature relationships.

7. References

1. Georgieva, E., Suta, M., & Burton, N. (2018). *HitPredict: Predicting BillboardHitsUsingSpotifyData*.StanfordMachineLearningPoster Session, Stanford University. Available from Stanford CS229 CS229MachineLearning
   .
2. Bertin-Mahieux,T.,Ellis,D.P.W.,Whitman,B.,&Lamere,P.(2011). *TheMillionSongDataset*.ISMIRConference.
3. Alghowinem,S.,etal.(2019).*SongHitPrediction:PredictingBillboard Hits Using Spotify Data*. Available from arXiv ar5iv
   .
4. Bertin-Mahieux,T.,etal.(2018).*PredictingaSong'sPaththroughthe Billboard Hot 100*. Machine Learning and Music Special Session, Stanford University.
5. Spotify Web API. (2023). Accessed for obtaining audio feature data, includingdanceability,energy,loudness,tempo,andmore.Availableat Spotify Developer.
6. Billboard.(2023).*BillboardHot100Chart*.Forreferenceonchartentry criteria and charting definitions. Available at Billboard.

Links

- **GoogleColabNotebook**: billboardhot100_predictor.ipynb
- **Dataset**: https://www.kaggle.com/api/v1/datasets/download/theoverman/the-spotify-hit-predictor-dataset