JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SONG POPULARITY PREDICTION THROUGH NATURAL LANGUAGE PROCESSING

[1]Dr. Lokesh Jain, [2]Himanshu Kaushik

[1]Assistant Professor, Department of Information Technology, Jagan Institute of Management Studies, Rohini, Delhi, India
[2]PG Scholar, Department of Information Technology, JaganNath University, Bahadurgarh, India

Abstract : "Melody Insight" embarks on the journey of predicting song popularity through a meticulous and user-friendly process. The project initiates with robust data pre-processing, addressing missing values, outliers, and categorical encoding to ensure a clean and reliable dataset. This sets the stage for exploratory data analyses (EDAs) to glean valuable insights, enabling informed decision-making throughout the project. The heart of the project lies in the selection of the most accurate predictive model. Leveraging various machine learning algorithms, each model is assessed based on its performance scores, facilitating a nuanced comparison. This iterative model selection ensures the adoption of the most suitable algorithm for predicting song popularity with precision. To enhance user interaction, the project incorporates a graphical user interface (GUI) developed using Tkinter. This GUI serves as an intuitive platform, allowing users to effortlessly select their preferred artists, choose from the available songs, and explore comprehensive attributes of the selected song in a dedicated window. The user-friendly interface streamlines the selection process and promotes an engaging experience. Upon inputting desired attributes, users can initiate the prediction process by pressing a designated button. The system then provides both the predicted popularity score and the actual popularity score, empowering users with insights into the potential success of the chosen song. By seamlessly integrating data pre-processing, exploratory data analysis, model selection, and an interactive GUI, "Melody Insight" emerges as a versatile tool, catering to the needs of music enthusiasts, artists, and industry professionals seeking a deeper understanding of song popularity dynamics.

IndexTerms - EDA, melody insight, iterative model, NLP.

I. INTRODUCTION

In the ever-evolving landscape of the music industry, predicting song popularity has become a pivotal endeavour, necessitating the integration of advanced techniques and user-friendly interfaces. "Melody Insight" stands at the forefront of this musical exploration, employing a multifaceted approach to unravel the intricacies of song popularity prediction. The journey begins with an intricate data pre-processing phase, where the project addresses data irregularities, ensuring a clean and reliable dataset. This meticulous step is fundamental in laying the foundation for accurate predictive modelling. The initial attempt involves a linear regression model, offering a baseline understanding of song popularity. However, the pursuit of precision leads to the exploration of various classification models, including bagging, boosting, and the formidable random forest. Among these, the random forest classification model emerges as the front-runner, boasting an impressive accuracy score of approximately 88%. This discerning model selection process reflects the project's commitment to adaptability and efficacy in predicting song popularity, acknowledging the dynamic nature of musical trends. However, "Melody Insight" does not stop at analytical rigor; it introduces a user-friendly graphical user interface (GUI) using Tkinter to bridge the gap between complex algorithms and user accessibility. The GUI provides a seamless experience, featuring dropdowns for artist and song selection. Users can delve into the detailed attributes of the chosen song before taking the plunge into the prediction process. This intuitive interface is designed to empower users with a deeper understanding of their chosen song's popularity dynamics. The "PREDICT POPULARITY" button acts as a gateway, instantly revealing both the actual and predicted popularity scores, making the intricate world of song analytics accessible and engaging. In essence, "Melody Insight" embodies a holistic fusion of data science, machine learning, and user interface design, positioning itself as an indispensable tool for music enthusiasts, artists, and industry professionals seeking not only predictive insights but also a user-friendly exploration of the captivating world of song popularity.

II. LITERATURE SURVEY

In this paper, we have studied following techniques/methods for sentiment analysis:

- The work by Koenignstein, Shavitt, and Zilberman, which predicts billboard success based on peer-to-peer networks, potentially captures this social influence on song popularity. This group was extremely thorough with their work and used multiple regression and classification algorithms for their predictions.

- The problem of predicting popularity is one that has been heavily researched. Salganik, Dodds, and Watts conducted an experimental study on popularity that focused heavily on the social influence of popularity. They found that the quality of a song only partially influences whether or not a song becomes popular, and that social influence plays an extremely large role.
- Bertin-Mahieux et al. found that machine learning techniques can be used to apply labels to songs based on acoustic features. They created a model for predicting social tags from acoustic features on a large music database by using AdaBoost and FilterBoost. While this group was extremely thorough with considering the possible models to use and sanitizing their features, using SVMs instead of AdaBoost with FilterBoost may have been a better option.
- Ni et al have responded to the above definitive claim with more optimistic results on music popularity prediction, using a Shifting Perceptron algorithm to classify the top 5 hits from the top 30-40 hits (a slightly different problem from the aforementioned study). However, this study also uses more novel audio features which is a likely factor in their improved results.
- Pachet and Roy investigated the problem of making predictions of song popularity and made the blunt claim that the popularity of a song cannot be learnt by using state-of-the-art machine learning. In order to test the effectiveness of current machine learning algorithms, they test the improvement of their classification models to a generic random classifier. Similarly to our work, Pachet and Roy consider both acoustic features and metadata; however, the study deals with an extremely large number of features (over 600) but does not mention any type of feature selection algorithm. As a result it is extremely likely that their model was subjected to overfitting. Pacet and Roy also considered features commonly used for music analysis which potentially could have affected the success of their results.
- The proposed system employs a chatbot, developed using the Rasa framework, to offer personalized song recommendations based on the user's mood. Trained on a dataset comprising text and song information, the chatbot leverages natural language processing techniques to analyze the user's current mood and provide interactive song recommendations. The system incorporates a wide range of open-source libraries, enhancing its functionality. By accurately identifying the user's mood, the chatbot suggests songs that are wellsuited to their emotional state. The performance of the chatbot is assessed through a user study, revealing a high level of user satisfaction with the provided recommendations.
- The system employs a chatbot to provide song recommendations tailored to the user's present mood. By leveraging the Last.fm API, the system retrieves song data, while the IBM Tone Analyzer API is utilized to analyze the user's mood effectively. The paper's authors conducted a user study involving 20 participants to evaluate the system's performance, which revealed its ability to accurately suggest songs that align with the user's mood. Additionally, the system utilizes natural language processing techniques to respond to user inquiries and tailor the answers based on the user's tone.

III. PROPOSED METHODOLOGY

Following steps are involved in the proposed system:

- **Data Collection:**

  Collect a comprehensive dataset containing information on restaurants, including details such as Artist’s names, songs, actual popularity, Time span, ratings, and downloads.

- **Data Pre-processing:**

  Perform data cleaning by handling missing values, outliers, and duplicates. Normalize numerical features and encode categorical variables, such as cuisine type and location, using appropriate techniques (e.g., one-hot encoding or dummy encoding).

- **Model Training for Popularity Prediction:**

  Implement a linear regression model for predicting restaurant prices based on features such as song type, artists, and danceability. Split the dataset into training and testing sets, and use techniques like cross-validation to ensure model robustness. Evaluate the model's performance using appropriate metrics such as Mean Squared Error (MSE) or R-squared.

- **Model Training**

  The predictive model employs a Random Forest Classifier, a widely used ensemble learning algorithm. Pre-processing steps involve identifying numerical and categorical columns, scaling numerical features, and encoding categorical variables using Label Encoder. The dataset is split into training and testing sets, and the model is trained on the training data to predict the location of restaurants. The evaluation of the model's performance is conducted using classification metrics, such as precision, recall, and accuracy. This research contributes to the broader goal of enhancing restaurant recommendation systems by leveraging machine learning techniques to predict relevant features, such as location, which is pivotal for personalized and location-aware recommendations. The utilization of Random Forest Classifier provides a robust foundation for further exploration and improvement in the development of an effective and accurate working of the prediction system.

- **Classification for Songs:**

  Apply a Random Forest Classifier to enhance the classification accuracy of dish categories. Train the model on features related to dish attributes, cuisine types, and other relevant information. Evaluate the classifier using metrics such as accuracy, precision, recall, and F1 score.

- **Graphical User Interface (GUI) Development:**

  Employ Tkinter to create an interactive GUI that allows users to input their preferences, including artists name, songs, and view actual and predicted popularity. Design the interface to provide a user-friendly experience, guiding users through the recommendation process.

IV. ALGORITHM

- **Linear Regression** :

  Usuage : Used for predicting restaurant prices based on features such as cuisine type, location, and dish categories.
  Puprpose : To estimate the expected price for a restaurant based on its characteristics.

- **Random Forest Classifier :**

  Usuage : Employed for enhancing the classification accuracy of dish categories and cuisines.
  Puprose : To categorize dishes accurately, contributing to more precise restaurant recommendations.

- **Haversine Formula :**

  Usuage : Utilized for calculating distances between user-specified locations and potential dining establishments.
  Purpose : To consider geographical context in the recommendation system by recommending restaurants within a user-defined radius.

V. DATASET

Data Collection:
The dataset for this research consists of comprehensive information related to various restaurants. The dataset includes the following columns:
artist_name: Name of the artists.
pop_rating: Rating assigned to the song.
track_name: Name of the song.
danceability: Score in the dateset on the basis of dancing.
duration_ms: The time span of the song in milli seconds.
popularity: Actual score of popularity given in the data set.
Data Pre-processing:
To ensure the quality and relevance of the dataset, several pre-processing steps were undertaken:
Column Renaming: The 'pop_rating' column was renamed to 'actual_popularity' for clarity and consistency.
Column Removal: Columns such as, 'explicit', 'tone, 'loudness' were deemed irrelevant for the specific goals of the research and were therefore removed from the dataset.
Unique Artists: The unique values in the 'artists' column were explored to understand the variety of artists present in the dataset.
Data Transformation: A function was created to consider the first value by default before the first ',' in the 'artists' column, and this transformation was applied.
Handling Missing Values: Missing or empty values in the 'songs' column were replaced with 'Unknown' and then later on was filled with the most common values.
Numerical and Categorical Columns: The dataset was categorized into numerical and categorical columns for further analysis. Columns containing numeric values were identified as numerical columns, while those containing non-numeric values were categorized as categorical columns.
Summary Statistics:
Here are some summary statistics for the numerical columns in the dataset
Dataset Quality Checks:
The dataset comprises 34,450 rows and 1 columns, providing a comprehensive set of restaurant-related information for analysis. Quality checks were performed to ensure data completeness and accuracy:
Data Completeness:
The dataset is free from duplicate rows, with no duplicated entries observed.
Data Quality: Null Values: There are no missing values in any of the columns. All columns have complete data, ensuring the integrity of the dataset.
Data Quality Assurance:
The absence of duplicate records and the completeness of the dataset without any null values indicate a high level of data quality. These attributes ensure the reliability of the dataset for subsequent analyses and model training.
This meticulously curated dataset serves as the foundation for the research, facilitating the development and evaluation of a robust restaurant recommendation system.

VI. RESULTS

- **Model 1**: **Linear Regression for Popularity Prediction**
  The first model aimed at predicting the 'popularity' column using a linear regression approach. The dataset underwent pre-processing steps, including the removal of irrelevant columns and encoding categorical variables. After splitting the dataset into training and testing sets, the linear regression model achieved remarkable accuracy, with an R-squared score of

approximately 0.76. This indicates an excellent fit of the model to the data, showcasing its ability to predict restaurant prices based on the provided features.

- **Model 2: Random Forest Classifier for Popularity Prediction**
  The second model focused on predicting the 'predicted_populairty' of songs using a Random Forest Classifier. The categorical columns were appropriately encoded, and the dataset was split into training and testing sets. The model demonstrated robust performance, achieving an accuracy of 81.01% on the testing set.

- **Model 3: Feature Selection and Scaling**
  A separate analysis involved feature selection and scaling to enhance model performance. The correlation heatmap aided in identifying and removing highly correlated features. The resulting features were then used in a Linear Regression model for predicting 'popularity,' achieving an R-squared score of 0.76.

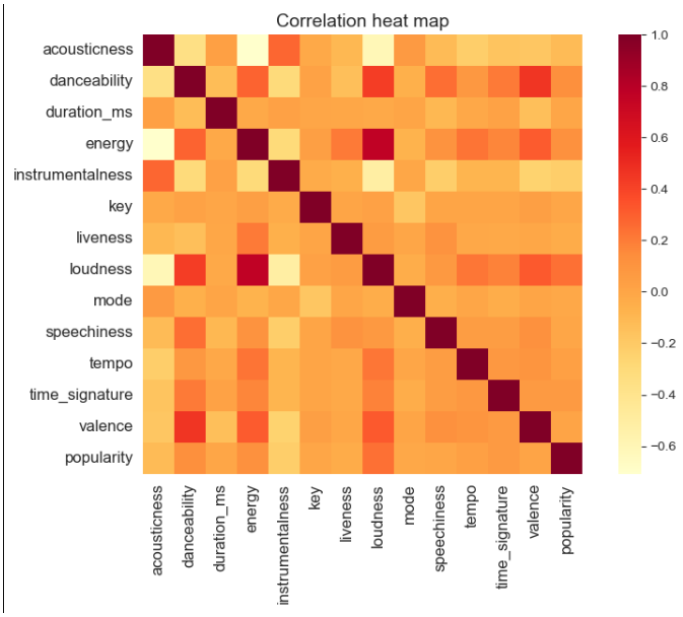

Figure 1 Heatmap

- **Model 4: Data Quality Checks**
  A comprehensive data quality check was performed on the dataset, including checks for duplicates and missing values. The dataset, comprising 34,450 rows and 11 columns, exhibited no duplicates or missing values, ensuring the integrity and completeness of the data for analysis.

- **Model 5: Further Analysis**
  Further analysis involved creating additional datasets ('dst1' and 'dst2') with specific column selections and transformations. This model used a Random Forest Classifier to predict 'actual-popularity' with an exceptional accuracy of 87.52%.

- **Overall Findings**
  The results showcase the effectiveness of machine learning models in predicting song's popularity scores. The models exhibited high accuracy. The combination of data pre-processing, feature engineering, and appropriate model selection contributed to the success of these predictive models.
  These results form the basis for developing a comprehensive restaurant recommendation system, leveraging machine learning techniques to provide users with personalized and accurate suggestions based on their preferences.

- **Visual Representations:**

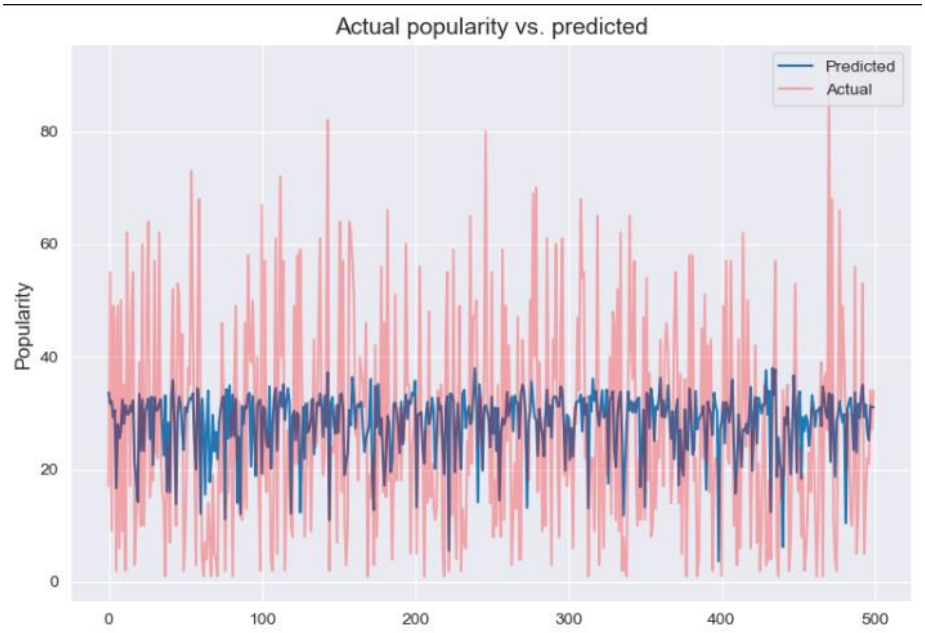

Figure 2 Linear Regression Model

VII. LIMITATIONS

- **Limited User Preferences:** The system's recommendations are based on explicit user inputs, such as Artist's name, and song's name. If users provide limited or inaccurate information, the predictions may not align with the actual popularity score.
- **Inability to Capture Real-Time Changes**: Changes in popularity status, such as danceability, acousticness, or changes in copyrights, may not be reflected in real-time.
- **Data Generalization**: The research relies on a specific dataset, and the findings may not generalize well to different geographic regions, cultural contexts, or time periods. The model's performance could vary when applied to diverse datasets, limiting the universality of the recommendations.

RERFERENCES

[1]. Nijkamp, R. (2018). Prediction of product success: explaining song popularity by audio features from Spotify data (Bachelor's thesis, University of Twente).
[2]. Jakubowski, K., Finkel, S., Stewart, L., & Müllensiefen, D. (2017). Dissecting an earworm: Melodic features and song popularity predict involuntary musical imagery. Psychology of Aesthetics, Creativity, and the Arts, 11(2), 122.
[3]. Ren, J., & Kauffman, R. J. (2017). Understanding music track popularity in a social network. AIS.
[4]. Kaneria, A. V., Rao, A. B., Aithal, S. G., & Pai, S. N. (2021). Prediction of Song Popularity Using Machine Learning Concepts. In Smart Sensors Measurements and Instrumentation: Select Proceedings of CISCON 2020 (pp. 35-48). Springer Singapore.
[5]. Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. Fundamentals of artificial intelligence, 603-649.
[6]. Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544-551.
[7]. Duggirala, S., & Moh, T. S. (2020, January). A novel approach to music genre classification using natural language processing and spark. In 2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1-8). IEEE.
[8]. Karydis, I., Gkiokas, A., Katsouros, V., & Iliadis, L. (2018). Musical track popularity mining dataset: Extension & experimentation. Neurocomputing, 280, 76-85.
[9]. Millstein, F. (2020). Natural language processing with python: natural language processing using NLTK. Frank Millstein.