Predicting Song Replay Likelihood Using Artificial Neural Networks

Internship Report

By

Rayesomo Timilehin Liberty

Student ID CA/FM3/2868

Submitted

To

CodeAlpha

Introduction

In the era of digital music streaming, understanding user listening patterns is crucial for enhancing user experience and improving music recommendation systems. This report details my approach to building a predictive model that estimates the likelihood of a user repeatedly listening to a song within a set timeframe. The project was conducted as part of my internship at CodeAlpha.

Dataset Overview

The dataset used for this project was sourced from Kaggle. It contained a large volume of user interaction data with songs, including features such as user ID, song ID, listening time, frequency of plays, song duration, and additional metadata. Given the dataset's size, Google Colab was used for model training to leverage its computational resources.

Data Preprocessing

To ensure the dataset was suitable for training a robust predictive model, the following preprocessing steps were performed:

1. **Loading the Data**: The dataset was imported into a Pandas DataFrame for exploratory analysis.
2. **Handling Missing Values**: Rows containing missing or inconsistent data were removed to maintain data integrity.
3. **Feature Selection and Engineering**: Relevant features were selected, and new features were engineered to enhance predictive power.
4. **Data Normalization**: Numerical features were scaled to improve the model’s convergence during training.
5. **Encoding Categorical Variables**: One-hot encoding and label encoding were applied where necessary.

Model Development

A deep learning-based Artificial Neural Network (ANN) was chosen to predict the likelihood of a user replaying a song. The model architecture consisted of the following layers:

A. **Input Layer**: Accepts the processed features as input.

B. **Hidden Layers**: Several fully connected layers with ReLU activation functions were used to capture complex relationships in the data.

C. **Output Layer**: A sigmoid activation function was used to output the probability of a song being replayed.

Model Training and Evaluation

i. The dataset was split into training and testing sets (80% training, 20% testing).

ii. The Adam optimizer and binary cross-entropy loss function were used to optimize the model.

iii. Early stopping and dropout techniques were implemented to prevent overfitting.

iv. The model was evaluated using accuracy, precision, recall, and the AUC-ROC curve to assess its predictive performance.

Challenges and Solutions

- **Large Dataset Handling**: Google Colab's GPU support was utilized for efficient training.
- **Imbalanced Data**: Oversampling techniques were applied to balance replay and non-replay cases.
- **Hyperparameter Tuning**: Various configurations of learning rates, batch sizes, and hidden layers were experimented with to improve performance.

Conclusion and Future Work

This project successfully developed a predictive model capable of estimating the likelihood of a user replaying a song. The insights derived from this model can be leveraged by music streaming platforms to enhance recommendation algorithms. Future improvements could involve incorporating additional user demographic data and exploring transformer-based deep learning models for improved accuracy.

I extend my gratitude to CodeAlpha for the opportunity to work on this project and to Kaggle for providing the dataset.