Development of Predictive Models for Spotify Track Success without Using sklearn

1. Introduction

The project embarked on a challenging yet innovative journey to develop predictive models that forecast the success metrics of Spotify tracks. Utilizing a dataset inclusive of various attributes like artist details, release dates, playlist appearances, streaming numbers, and musical features (danceability, energy, valence), the aim was to build effective models from scratch, explicitly avoiding the convenience of the sklearn library.

2. Objective

The primary goal was to construct predictive models capable of accurately forecasting Spotify track success, leveraging manual implementations of algorithms to deepen the understanding of machine learning fundamentals and enhance customization capabilities.

3. Methodology

The methodology section underscores the project's commitment to a manual approach:

Data Collection: Sourcing a dataset that encompasses comprehensive attributes of Spotify tracks up to 2023.

Data Cleaning and Preparation: Handling missing values, outlier detection, and normalization of features without the automated tools provided by sklearn.

Feature Engineering: Manually selecting and transforming features to better represent the underlying patterns influencing track success.

4. Model Development

Two models were developed, emphasizing the project’s innovative approach:

Linear Regression (Manual Implementation): Constructed without sklearn, this model’s development included manual coding of regression algorithms, calculation of coefficients through gradient descent, and evaluation using custom functions to compute MSE and R^2 Score.

XGBoost (Manual Setup): Despite its complexity, the model was built from the ground up, focusing on ensemble learning principles, decision tree construction, and boosting techniques to optimize performance.

5. Preprocessing of Data

In-depth data preprocessing involved:

Normalization: Manually scaling feature values to a standard range to ensure equal importance during model training.

Encoding: Converting categorical variables into numerical formats through manual encoding techniques, laying a foundation for model ingestion.

6. Implementation of Algorithm

The manual implementation of algorithms revealed:

Linear Regression Results: Achieved a MSE of 0.24588130147883713, indicating the model’s efficiency in predicting Spotify track success.

XGBoost Results: Presented a Mean MSE of 0.3487515549821321, highlighting its performance variability and the challenges of manual ensemble model construction.

7. Bias-Variance Trade-off Analysis

This section delves into the nuanced balance between model simplicity and complexity, highlighting the Linear Regression model’s tendency towards high bias and the XGBoost model’s variance, providing insights into their potential underfitting and overfitting behaviors.

8. Model Selection and Generalization

Based on empirical data, Linear Regression was chosen for its favorable bias-variance trade-off and generalization potential, especially given the manual development context which excludes sklearn's optimizations.

9. Overfitting and Underfitting Considerations

Detailed strategies to combat overfitting in the complex XGBoost model and underfitting in the simpler Linear Regression model are discussed, emphasizing the importance of cross-validation and robust pattern recognition in manual modeling.

10. Conclusion

The report concludes with a reflection on the journey of manually developing predictive models for Spotify track success, emphasizing the Linear Regression model’s balance between accuracy and generalizability. This project not only achieved its objective of forecasting track success but also demonstrated the feasibility and educational value of building predictive models without relying on sklearn.