Milestone 1 Report (ML)

Team ID: SC_10

Team Names & IDs:

عبدالفتاح محمد حسين حسين (2021170316)

عبدالمنعم محمد عادل أحمد (2021170325)

ايات محمد عبدالعزيز عبدالشافي (20201700163)

منة الله محمد علي (20201701145)

سعد محمود سعد العزازي (2021170230)

سامح خليل ابراهيم خليل (2021170228)

Preprocessing Techniques:

- Dropping Unnecessary Columns: dropped columns like 'Song', 'Album', 'Album Release Date', etc., which are not likely to directly influence the popularity prediction. This was done using the DataFrame.drop() function.

- One-Hot Encoding: is a technique used to convert categorical columns into binary indicators. we've applied one-hot encoding using pd.get_dummies().

- Feature Scaling: is used to ensure that all features have the same scale. Here, we've used (z-score normalization) to scale the features using StandardScaler from Scikit-Learn. This was implemented in the feature_scaling() function.

- Missing Values Handling: Dropped rows with missing values using dropna().

Dataset Analysis:

- Correlation Analysis: Explored the correlation between numerical features to identify relationships. Generated a heatmap to visualize the absolute correlation values.

- Feature Selection: Used SelectKBest with f_regression to select the top k features that have the strongest linear relationship with the target variable 'Popularity'.

Regression Techniques:

- Linear Regression: Trained a Linear Regression model using sklearn's linear_model.LinearRegression().
- Random Forest Regression: Employed a Random Forest Regression model using sklearn's RandomForestRegressor().

Differences between Models:

Linear Regression vs. Random Forest:

- Linear Regression: Achieved an MSE of 0.48718422 and an R2 score of 0.5231719727 and accuracy of 52%.
- Random Forest Regression: Initially, achieved an MSE of 0.464544771 and an R2 score of 0.54533017. After hyperparameter tuning, achieved an improved score with an MSE of 0.46454477 and an R2 score of 0.56306257 and final accuracy of 56%.

Features Used/Discarded For Each Model:

('Hot100 Ranking Year', 'Hot100 Rank', 'Song Length(ms)', 'Acousticness', 'Danceability', 'Energy', 'Instrumentalness', 'Liveness', 'Loudness', 'Speechiness', 'Mode', 'Time Signature', ..etc).

The Size Of The Validation, Test, Train Sets:

Train-Test Split: Before training the models, the dataset is split into training and testing sets using train_test_split() from Scikit-Learn. This ensures that the model's performance can be evaluated on unseen data. (test_size=20%,train_size=80%), we didn’t make validation.

Further Techniques Used For Improvement:

- Hyperparameter Tuning: Conducted hyperparameter tuning for the Random Forest Regression model using GridSearchCV to optimize model performance.

Conclusion About Project (Phase 1):

In this phase of the project, we preprocessed the dataset by cleaning, encoding, and scaling the features. We explored the dataset through correlation analysis and feature selection. Two regression techniques, Linear Regression and Random Forest Regression, were employed to predict song popularity. While both models showed reasonable performance, Random Forest Regression outperformed Linear Regression after hyperparameter tuning. The feature selection process helped in identifying the most relevant features for prediction. Further analysis and model refinement will be carried out in subsequent phases.