CSCI 6364 Machine Learning - Assignment 1

Abde Manaaf Ghadiali – G29583342

Disclaimer – Much of the Information for Question 1 has been excluded from the report due to the report's length restriction. Please refer to the Jupyter Notebook which is present along with this report in the zip file.

Question 1: Machine Learning from Scratch: Kaggle Most Streamed Spotify Songs 2023

Objective

Our Aim with this Assignment is to implement a machine learning algorithm from scratch (using first principal functions) and predict the number of Streams using the "***Most Streamed Spotify Songs 2023***" data using various features such as song's popularity across different music platforms, artist involvement and attributes, audio features, temporal features, etc.

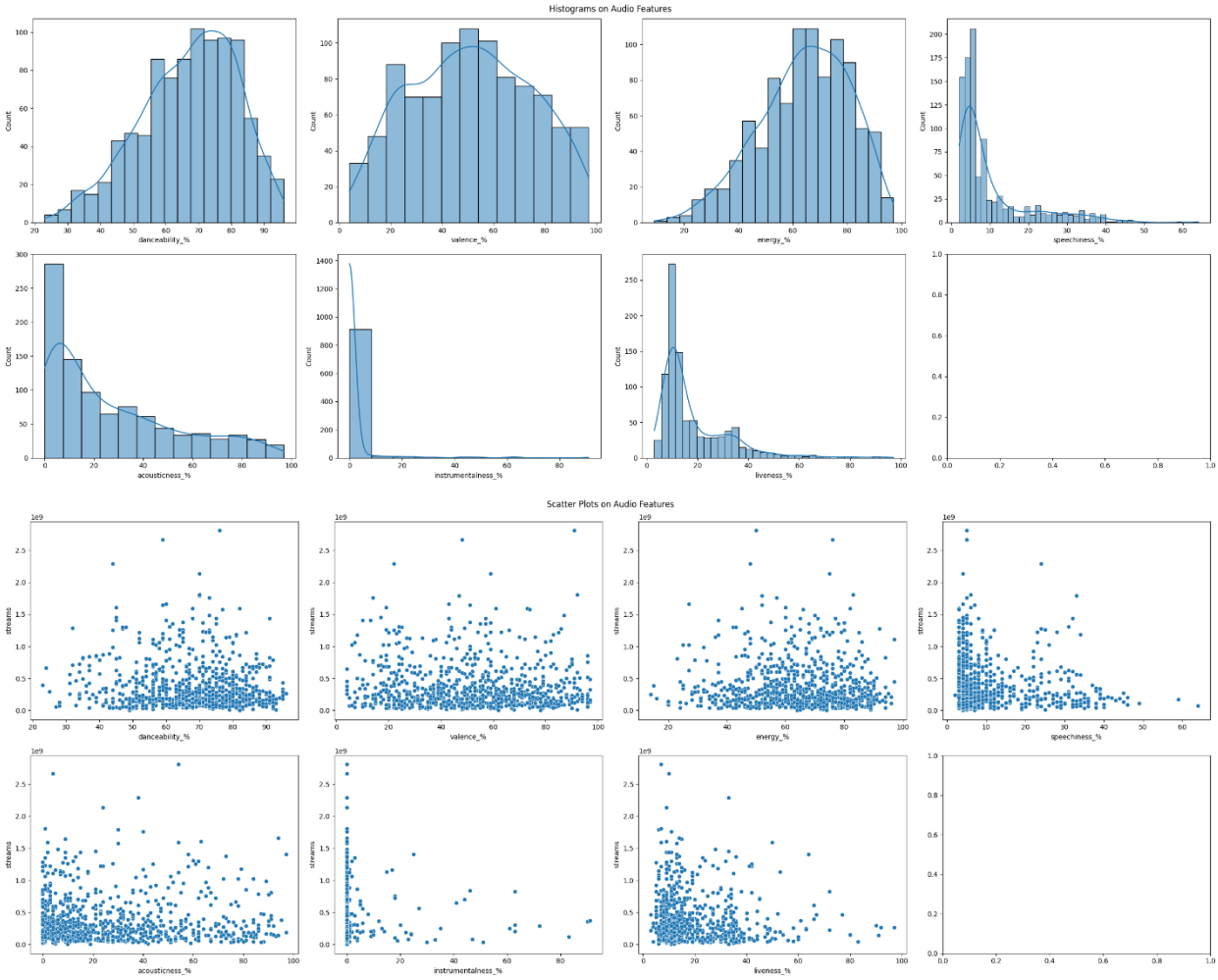
Data Source: https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023.

Data Cleaning and Preprocessing

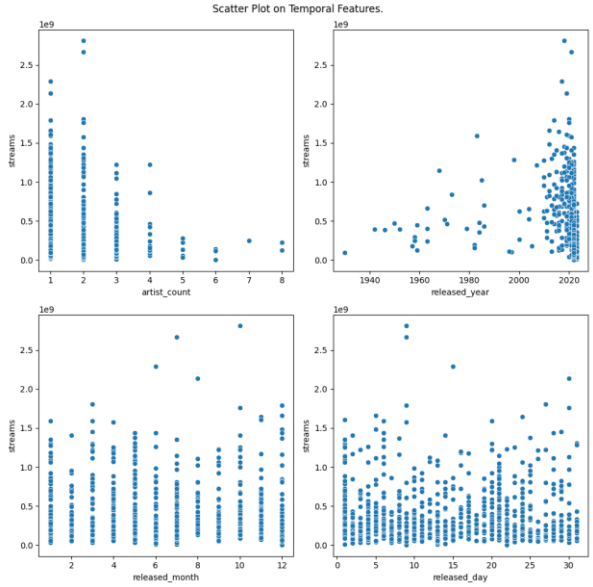
1. The 'Stream Feature' variable within the raw dataset was found to contain non-numeric values due to an erroneous entry. To address this issue, the erroneous values were coerced to NaN (Not a Number) values, and the corresponding rows were subsequently removed from the dataset.
2. The variables 'in_deezer_playlists' and 'in_shazam_charts' contained numeric values with commas, resulting in the dtype being categorized as 'object'. To rectify this, the commas were replaced with empty strings, and the entire column was converted to a numeric format.
3. In consideration of data relevance, redundant features such as 'track_name' and 'artist(s)_name' were excluded from further analysis.
4. Two variables, 'key' and 'in_shazam_charts', exhibited missing values. After thorough examination of the dataset and analysis of related features, it was determined that if a track did not appear in the Shazam charts, the corresponding value should be filled with '0' to represent absence.
5. Addressing missing values in the 'key' feature necessitated a distinct approach. NaN values were replaced with the mode of the keys associated with that artist, provided the artist had at least one track with a non-null key entry. Alternatively, NaN values were imputed with 'NA' (string) where no such information was available.
6. Among the variables, 'key' and 'mode' are identified as categorical features. Dummy Variables (One-Hot Encoded Features) we created, and the categorical features were dropped.

Exploratory Data Analysis (EDA)

1. The raw dataset comprises 24 features and contains a total of 953 records. The target feature under investigation is 'streams'.
2. Utilizing the Pandas 'describe' function, comprehensive statistical summaries such as mean, median, minimum, maximum, etc., are obtained for each feature. Notably, analysis reveals that 'track_name' is not inherently unique, suggesting that a composite key combining 'track_name' and 'artist(s)_name' is warranted.
3. Additionally, it is observed that Taylor Swift boasts the highest count of tracks among solo artists.
4. The dataset encompasses audio-related attributes, including 'danceability_%', 'valence_%', 'energy_%', 'acousticness_%', 'instrumentalness_%', 'liveness_%', and 'speechiness_%'.
5. Initial examination of the audio features reveals that 'instrumentalness_%' predominantly registers values close to 0. By conducting T-Statistic analysis and deriving p-values for all audio attributes, it is discerned that 'danceability_%' and 'speechiness_%' exhibit significance concerning the target variable, 'streams'. Consequently, the removal of non-significant audio features is recommended.

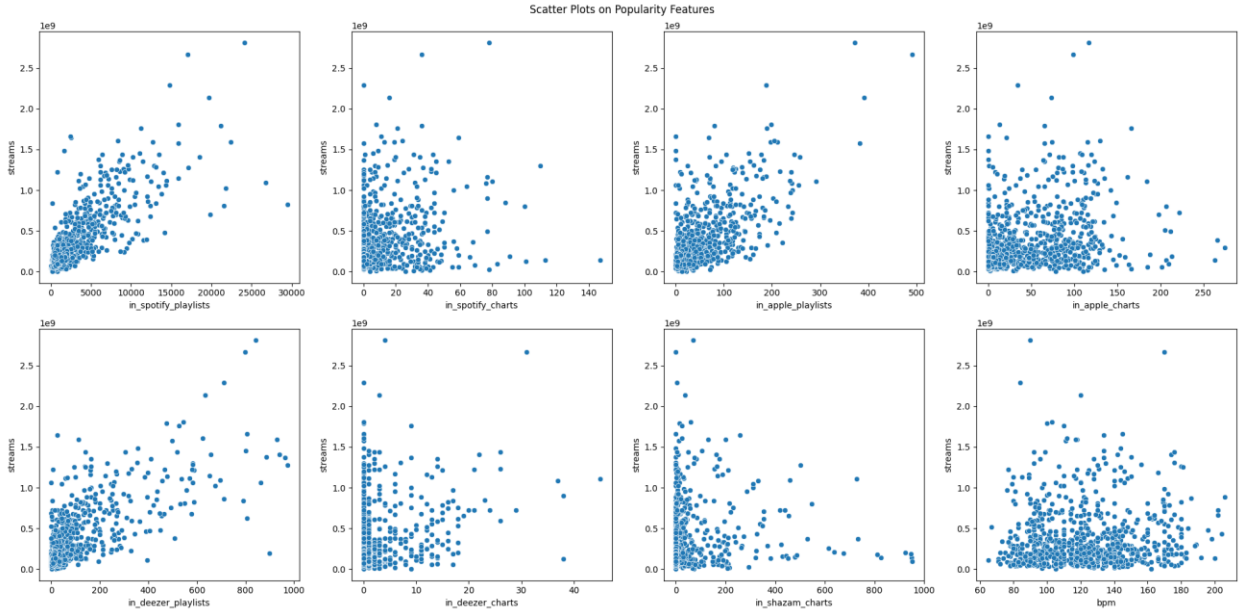


6. Features such as 'released_year', 'released_month', and 'released_day' are temporal in nature. Notably, the p-value associated with 'released_year' indicates its significance. Consequently, a novel feature, 'new_releases_2010', is proposed to categorize songs based on their release year relative to 2010. The determination of 2010 as the threshold is guided by insights gleaned from the scatter plot.

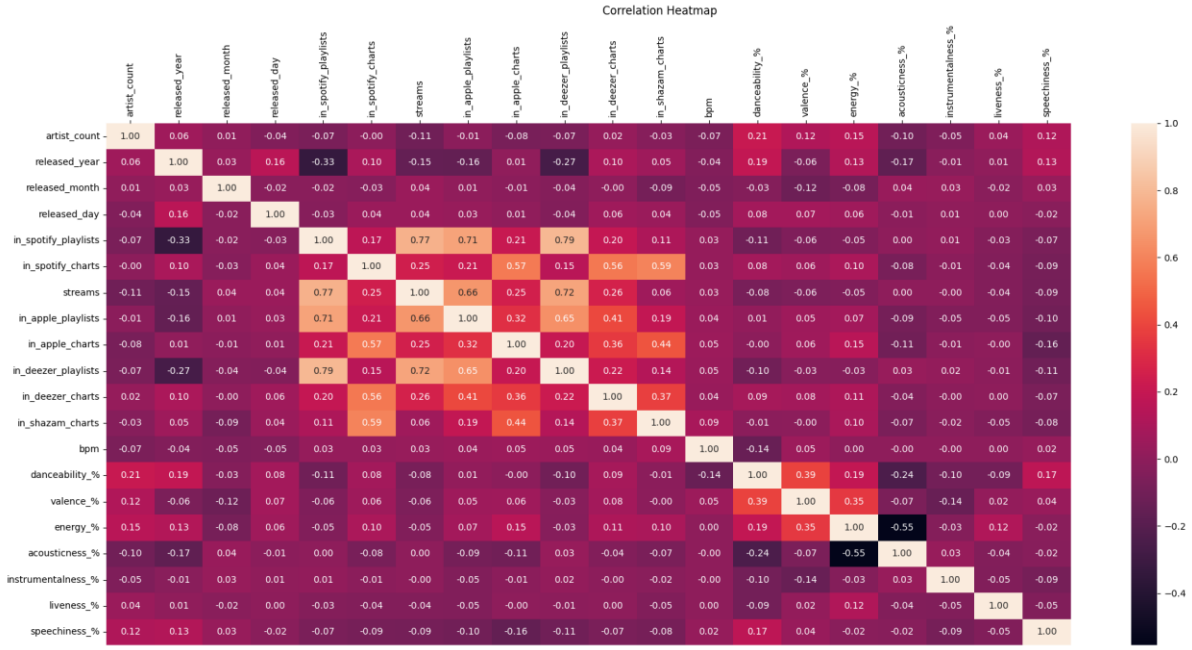


7. The 'artist_count' feature denotes the total number of artists associated with each track. The p-value associated with artist_count' indicates its significance.

8. Given the categorical nature of 'key' and 'mode' features, dummy variables are generated through One-Hot Encoding to facilitate analysis. Subsequent application of T-Statistic enables the calculation of p-values. Evaluation of these p-values suggests that 'key_A' and 'key_C#' demonstrate notable relevance.
9. All the popularity features demonstrate a meaningful association with the target variable. Through comprehensive analysis, it is evident that metrics such as 'in_spotify_playlists', 'in_spotify_charts', 'in_apple_playlists', 'in_apple_charts', 'in_deezer_playlists', 'in_deezer_charts', and 'in_shazam_charts' exhibit varying degrees of correlation or influence on the target variable.



10. This visualization elucidates the interrelationships among features by illustrating their correlations. It offers insights into which features are correlated with one another, aiding in the identification of potential multicollinearity issues, and informing feature selection processes.



Model Implementation and Evaluation

Model: Linear Regression (Supervised)
Optimizer: Gradient Descent with Loss Function as MSE
Evaluation Metric: R-Squared
Regularization: L2

The model construction process commenced with data shuffling and partitioning into training and testing subsets. To ensure uniformity in feature scales and facilitate optimal performance, the data underwent scaling using the min-max scaler, thereby constraining values within a positive range. Subsequently, a linear regression model incorporating L2 regularization with a learning rate of 0.1 was fitted to the scaled training data.

To ascertain the model's ability to generalize effectively, cross-validation was conducted. This technique involves partitioning the dataset into complementary subsets, fitting the model on a portion of the data, and evaluating its performance on the remaining portion. By iteratively repeating this process across different partitions, the model's robustness and generalization capabilities can be assessed. Following cross-validation, the model's performance was evaluated on a holdout set that had not been utilized during model training or cross-validation. This step is crucial for providing an unbiased estimate of the model's predictive performance on unseen data.

Below are the key findings from the evaluation of the model on both the training and testing datasets:

Training Data Performance:

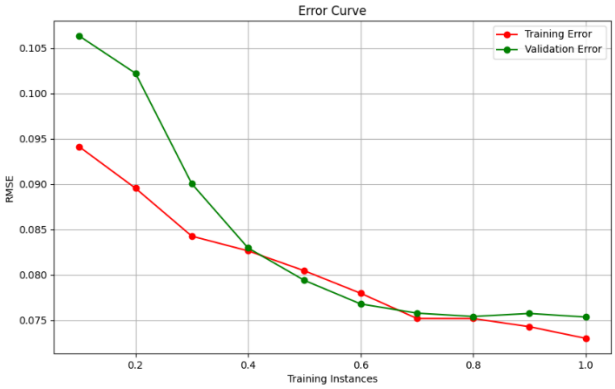
- Mean Squared Error (MSE): 0.00533
- R-squared (R2) Score: 0.8006
- Root Mean Squared Error (RMSE): 0.07299

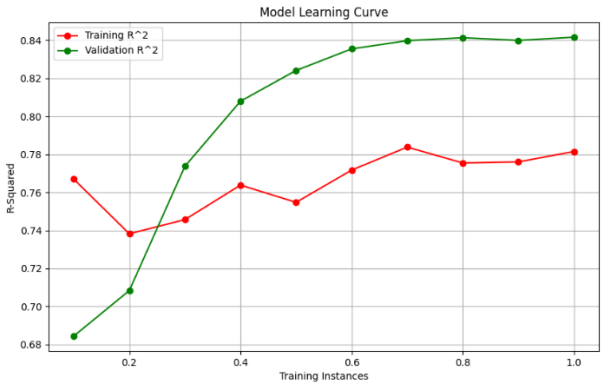
Validation Data Performance:

- Mean Squared Error (MSE): 0.00567
- R-squared (R2) Score: 0.84172
- Root Mean Squared Error (RMSE): 0.07530

Testing Data Performance:

- Mean Squared Error (MSE): 0.00521
- R-squared (R2) Score: 0.73127
- Root Mean Squared Error (RMSE): 0.07218





We may confidently assert that our model exhibits favorable performance regarding our Error (Loss) function. Notably, the consistency observed in the loss values across the training, validation, and testing datasets indicates a balanced bias-variance tradeoff at an optimal level. The selection of our final model was meticulously guided by iterative experimentation to ascertain the optimal learning rate and number of iterations. Through this process, we identified the model configuration that yielded the most favorable loss values for both training and validation datasets.

The outcome metrics and accompanying graphical representations collectively attest to the efficiency with which our model generalizes the data. Furthermore, an analysis of the performance metrics reveals minimal evidence of overfitting, as indicated by the negligible disparity between training and testing errors. This suggests that our model achieves a satisfactory level of robustness and generalization, underscoring its effectiveness in capturing the underlying patterns inherent in the data.

Question 2: Bias-Variance Tradeoff

A. GaussianNB attains optimal bias and variance equilibrium at approximately 1100 training instances, while SVC achieves this balance around 750 training instances. This inference is drawn from the convergence of training and cross-validation scores depicted in the graph, indicating the model's ability to generalize effectively. Further expansion of training beyond these thresholds is deemed superfluous.

B. Model Operating Regimes:
   a. With a dataset size of 250 data points, both GaussianNB and SVC exhibit overfitting tendencies, characterized by high variance due to the limited number of training instances.
   b. With 1000 or more data points, GaussianNB and SVC demonstrate optimal bias and variance characteristics, although SVC outperforms GaussianNB, owing to improved generalization capabilities.

C. Mitigation Strategies for High Bias and High Variance:
   a. High Bias (Underfitting): Models exhibiting high bias struggle to capture underlying data patterns effectively. Strategies to address this include augmenting model complexity by introducing additional features, increasing model capacity, and reducing regularization.
   b. High Variance (Overfitting): Models with high variance are overly sensitive to training data fluctuations, often memorizing noise rather than genuine patterns. Techniques to mitigate overfitting entail reducing model complexity by limiting features, decreasing model capacity, and augmenting regularization.

D. The incremental addition of data may not necessarily enhance model performance beyond its optimal bias-variance threshold. Instead, it may lead to increased training time and the risk of overfitting, as the model has already achieved a sufficient level of generalization.

E. Training and Validation Score Plot for Underfitting Model: