# Regression Assignment

## Kruti Patel

## Mohawk College

## CSIT-Software Development-559

## Professor: Sara Zendehboodi

## 27/02/2024

# Data Description

Data Set Name: Song Popularity Dataset

##### Source:

Humans have greatly associated themselves with Songs & Music. It can improve mood, decrease pain and anxiety, and facilitate opportunities for emotional expression. Research suggests that music can benefit our physical and mental health in numerous ways.

Lately, multiple studies have been carried out to understand songs & it is popularity based on certain factors. Such song samples are broken down & their parameters are recorded to tabulate. Predicting the Song Popularity is the main aim.

Description of Features:

The dataset consists of 15 features, which are used to predict the Song Popularity. The features include both numerical and categorical variables. The features are:

* **song\_name:**  The name or title of the song. It is an object data type.
* **song\_popularity:** An integer value representing the popularity of the song.
* **song\_duration\_ms**: The duration of the song in milliseconds
* **acousticness:** A measure of the acoustic quality of the song.
* **danceability:** A measure of how suitable the song is for dancing based on musical elements.
* **energy:** Represents the intensity and activity of the song.
* **instrumentalness:** Indicates the likelihood that the song is instrumental.
* **key:** Represents the key to the song, which is a musical attribute.
* **liveness:** Indicates the presence of an audience in the recording.
* **loudness:** The overall loudness of the song, measured in decibels (dB).
* **audio\_mode:** Represents the modality of the track.
* **speechiness:** Measures the presence of spoken words in the song.
* **tempo:** The speed or pace of the song, measured in beats per minute (BPM).
* **time\_signature:** Represents the number of beats in each bar of music.
* **audio\_valence:** Describes the musical positiveness of the song.

The dataset contains a mix of float64 (decimal numbers), int64 (integer numbers), and object (text) data types, providing a diverse set of features to analyze and model song popularity**.**

Description of Regression Task:

##### The regression task is to predict the predict the Song Popularity. (between 0-100) based on the given features.

##### Statistics:

Number of features: 15

No of examples: 18835

Range of each feature:

|  |  |
| --- | --- |
| feature | Range |
| song\_name | [0,18834] |
| song\_popularity | [0, 100] |
| song\_duration\_ms | [12.0k, 1.90m] |
| acousticness | [0, 1] |
| danceability | [0,0.99] |
| energy | [0,1] |
| instrumentalness | [0,1] |
| key | [0, 11] |
| liveness | [0.01,0.99] |
| loudness | [-38.8,1.58] |
| audio\_mode | [0,1] |
| speechiness | [0.0294, 0.0302] |
| TEMPO | [91.49,167.06] |
| time\_signature | [3,4] |
| audio\_valence | [0.474,0.581] |

*Table1*

# Note:

The "song\_name" feature had to be omitted from the dataset for regression analysis due to its data type being non-numeric (object). Regression models require numerical input features to make predictions, and since song names are text-based, they cannot be directly used for this purpose. The range varies for each feature, and they have different scales. For example, SONG\_DURATION\_MS ranges from 12.0k to 1.90m, while LOUDNESS ranges from -38.8to 1.58. Due to this reason, I perform normalization before using this dataset in regression models.

**No of Features:**14

**No of examples**:18835

# The Tests

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | DESCRIPTION | Mean MSE | | Min MSE | | MAX MSE |
| LINEAR | N.A. | 457.336 | 442.961 | | 483.795 | |
| POLYNOMIAL | Order=2 | 456.958 | 397.074 | | 592.643 | |
| K-NN | n\_neighbors=28, weights='distance' | 327.010 | 289.967 | | 349.354 | |
| DECISION TREE | max\_depth=10, min\_samples\_leaf=6, min\_samples\_split=200 | 452.397 | 434.885 | | 470.01 | |
| Support Vector Machines | kernel='linear', C=50, gamma=0.001, epsilon=0.1 | 467.350 | 453.032 | | 495.483 | |
| rANDOM FOREST | n\_estimators=300, max\_depth=10, max\_features=12 | 369.380 | 360.717 | | 389.914 | |

*Table2*

# Discussion:

Linear Regressor:

The Linear Regression model does not have any hyperparameters to tune. It models the relationship between the independent and dependent variables as a linear function. The Linear Regressor has the highest average MSE (457.336), indicating that it may not be the best choice for this dataset. This could be due to the simplicity of the model, which may not accurately capture the underlying patterns in the data.

Polynomial Regressor:

The Polynomial Regression model has an order parameter of 2, which means that it models the relationship between the independent and dependent variables as a polynomial function of degree 2. The Polynomial Regressor has an average MSE (456.958) that is slightly lower than the Linear Regressor, but it has a much higher maximum MSE (592.643), indicating potential overfitting. This could be due to the complexity of the model, which may not be necessary for this dataset.

KNN Regressor:

The KNN Regressor model has the hyperparameters n\_neighbors=28 and weights='distance'. This means that it models the relationship between the independent and dependent variables based on the 28 nearest neighbors, where the prediction is weighted by the inverse of the distance to the neighbor. The KNN Regressor has the lowest average MSE (327.010), indicating that it is the best choice for this dataset. This could be due to its simplicity and ability to capture the underlying patterns in the data.

Decision Tree Regressor:

The Decision Tree Regression model has the hyperparameters max\_depth=10, min\_samples\_leaf=6, and min\_samples\_split=200. This means that it models the relationship between the independent and dependent variables as a decision tree with a maximum depth of 10, where each leaf node has a minimum of 6 samples and each internal node has a minimum of 200 samples. The Decision Tree Regressor has an average MSE (452.397) that is slightly lower than the Linear Regressor and Polynomial Regressor, but it has a higher maximum MSE (470.01), indicating potential overfitting. This could be due to the complexity of the model, which may not be necessary for this dataset.

Support Vector Regressor:

The Support Vector Regressor model has the hyperparameters kernel='linear', C=50, gamma=0.001, and epsilon=0.1. This means that it models the relationship between the independent and dependent variables as a hyperplane in a high-dimensional space, where the kernel function is linear, the regularization parameter C is 50, the gamma kernel is 0.001, and the tolerance for the epsilon is 0.1. The Support Vector Regressor has an average MSE (467.350) that is higher than the KNN Regressor, indicating that it may not be the best choice for this dataset. This could be due to the complexity of the model, which may not accurately capture the underlying patterns in the data.

Random Forest Regressor:

The Random Forest Regressor model has the hyperparameters n\_estimators=300, max\_depth=10, and max\_features=12. This means that it models the relationship between the independent and dependent variables as a collection of 300 decision trees, where each tree has a maximum depth of 10 and each node splits on at most 12 features. The Random Forest Regressor has an average MSE (369.380) that is lower than the Linear Regressor, Polynomial Regressor and Support Vector Regression, but it is higher than the KNN Regressor. This could be due to the complexity of the model, which may not be necessary for this dataset.

In conclusion, based on the provided hyperparameters, the KNN Regressor is the best choice for this dataset, as it has the lowest average MSE (327.010). The Linear Regressor, Polynomial Regressor, and Support Vector Regressor have higher average MSEs, indicating that they may not be the best choice for this dataset. The Decision Tree Regressor and Random Forest Regressor have lower average MSEs, but they have higher maximum MSEs, indicating potential overfitting. Therefore, I would recommend using the **KNN Regressor** for this dataset.

## References

Lee, T. T. T. C. a. M. (2023, December 23). *Chapter 7 Regression I: K-nearest neighbors |  Data Science*. https://datasciencebook.ca/regression1.html

Saxena, S. (2023, August 25). *A Beginner’s Guide to Random Forest Hyperparameter Tuning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/

Sethi, A. (2024, February 9). *Support Vector Regression tutorial for machine learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

*Song popularity Dataset*. (2022, January 7). Kaggle. https://www.kaggle.com/datasets/yasserh/song-popularity-dataset

Verma, N. (2023, November 2). An introduction to Support Vector Regression (SVR) in machine learning. *Medium*. https://medium.com/@nandiniverma78988/an-introduction-to-support-vector-regression-svr-in-machine-learning-681d541a829a