Project Introduction

Name: Prateek Singh

Roll No: 2301156

Project Name: Music Prediction using Machine Learning

Report: Music Prediction Using ML

Algorithms Details

Description:

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's 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.

The project is simple yet challenging, to predict the song popularity based on energy, acoustics, instumentalness, liveness, dancibility, etc. The dataset is large & it's complexity arises due to the fact that it has strong multicollinearity. Can you overcome these obstacles & build a decent predictive model?

Acknowledgement:

The dataset is referred from Kaggle.

Objective:

- Understand the Dataset & cleanup (if required).
- Build Regression models to predict the music popularity.
- Also evaluate the models & compare their respective scores like R2, RMSE, etc.

Interpretation

We aim to solve the problem statement by creating a plan of action, Here are some of the necessary steps:

1. Data Exploration
2. Exploratory Data Analysis (EDA)
3. Data Pre-processing
4. Feature Selection/Extraction
5. Predictive Modelling
6. Project Outcomes & Conclusion

1. Data Exploration

The stats seem to be fine, let us do further analysis on the Dataset

2. Exploratory Data Analysis (EDA)

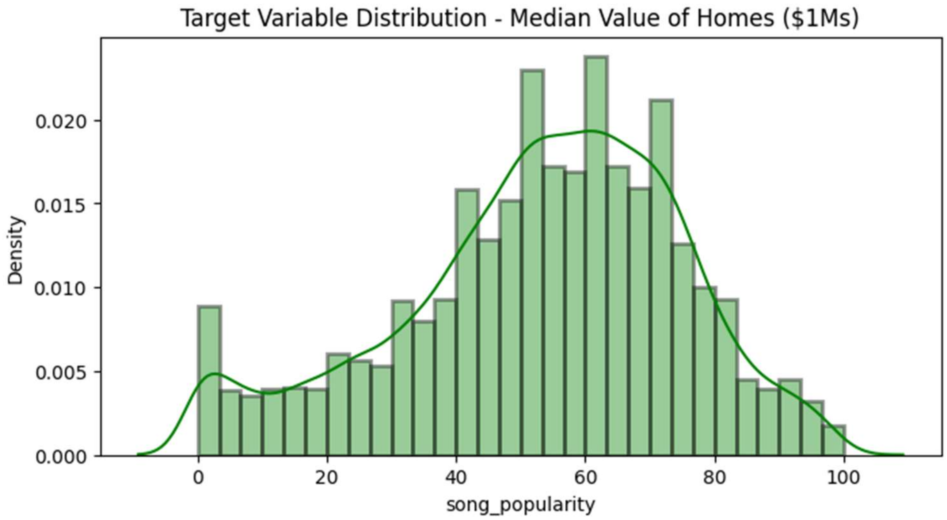

Figure 1: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

Inference: The Target Variable seems to be be normally distributed, averaging around 60 units.

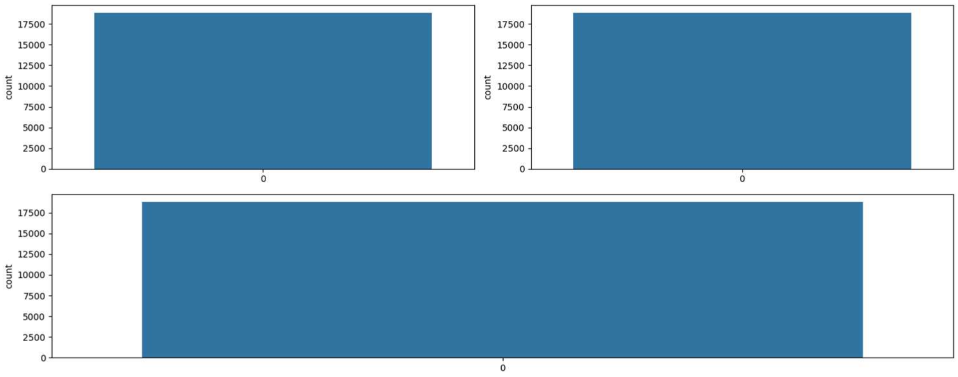

Figure 2: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

Inference: The categorical features reveal alot of information about the dataset.

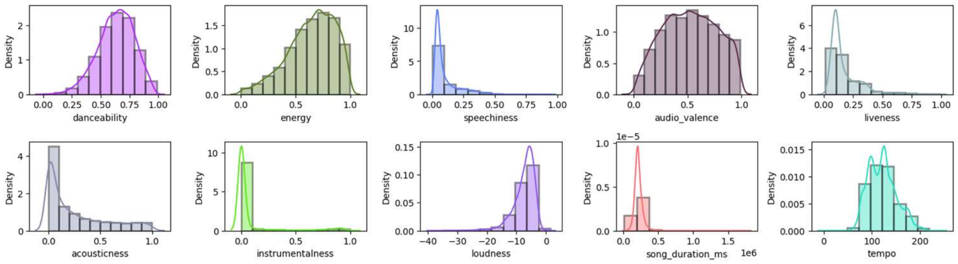

Figure 3: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

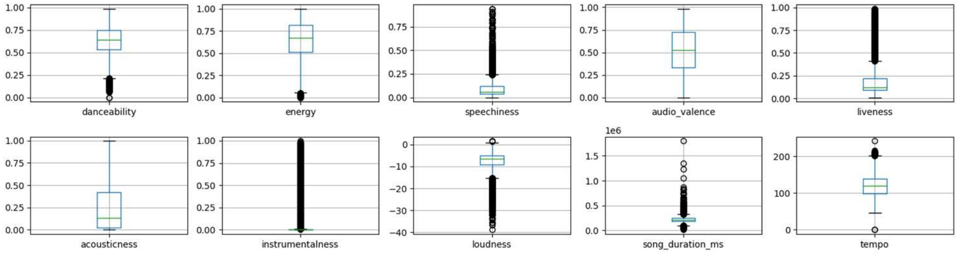

Figure 4: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

Inference: There seem to be some outliers. let us fix these in the upcoming section...

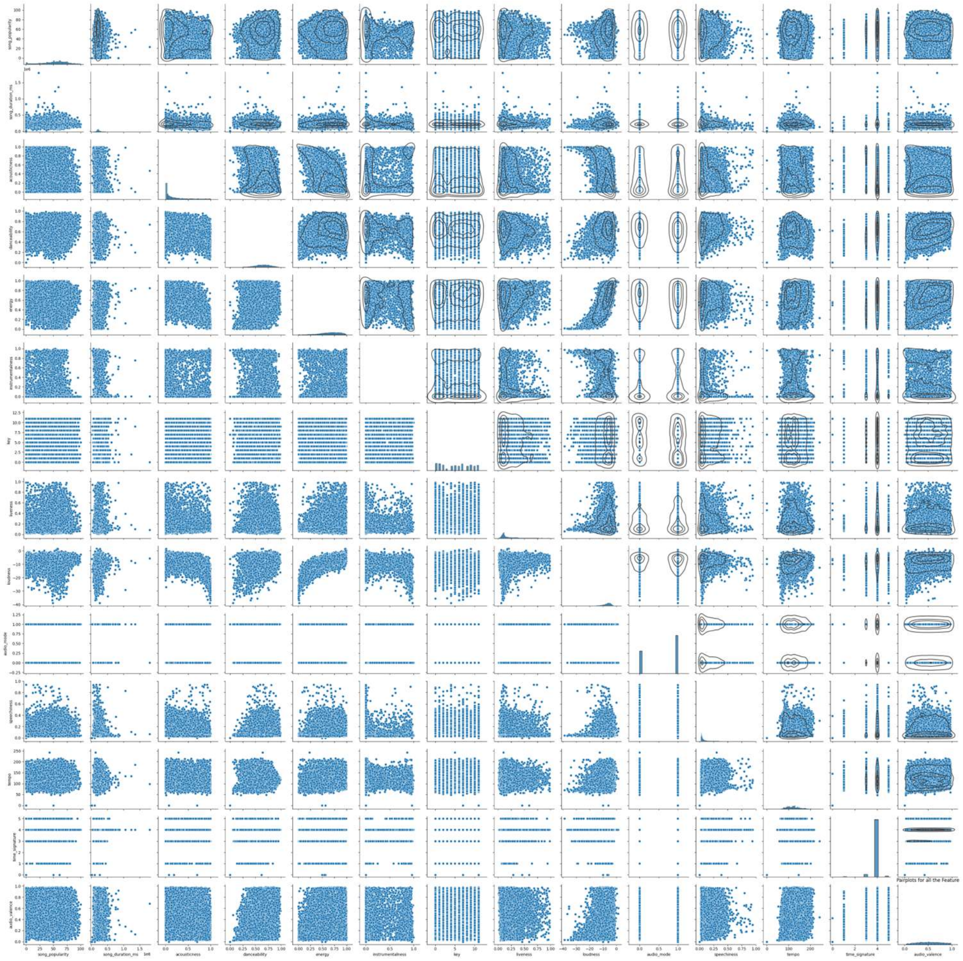

Figure 5: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

Inference: We can notice that some features have linear relationship, let us futher analyze the detect multicollinearity.

3. Data Preprocessing

The datset doesn't have any inconsistant values.

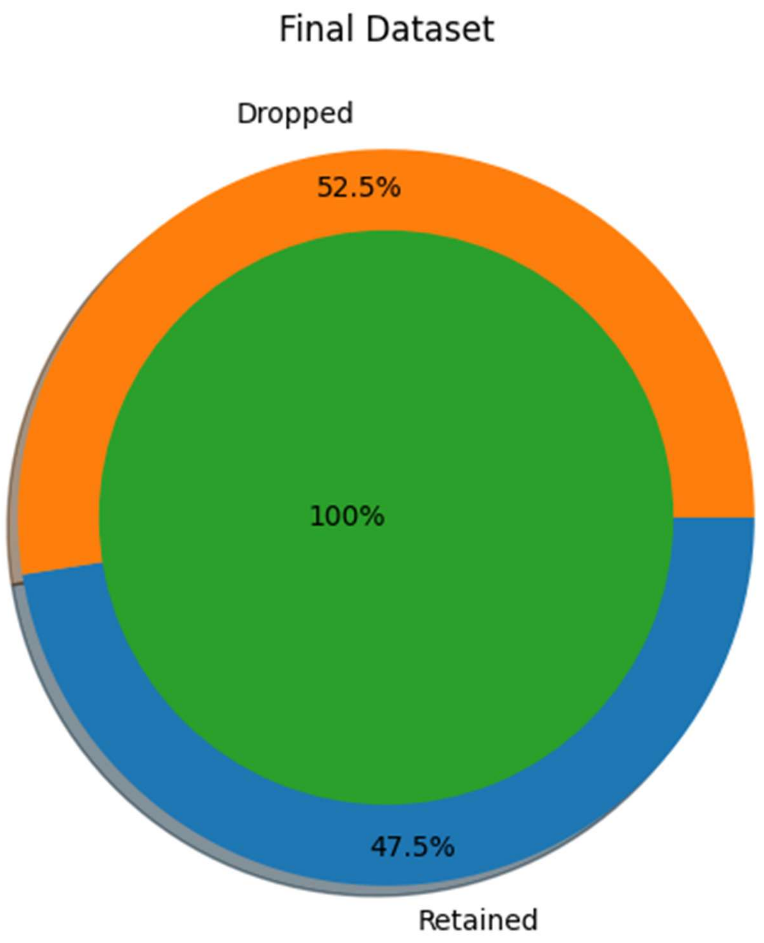

Figure 6: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

4. Feature Selection/Extraction

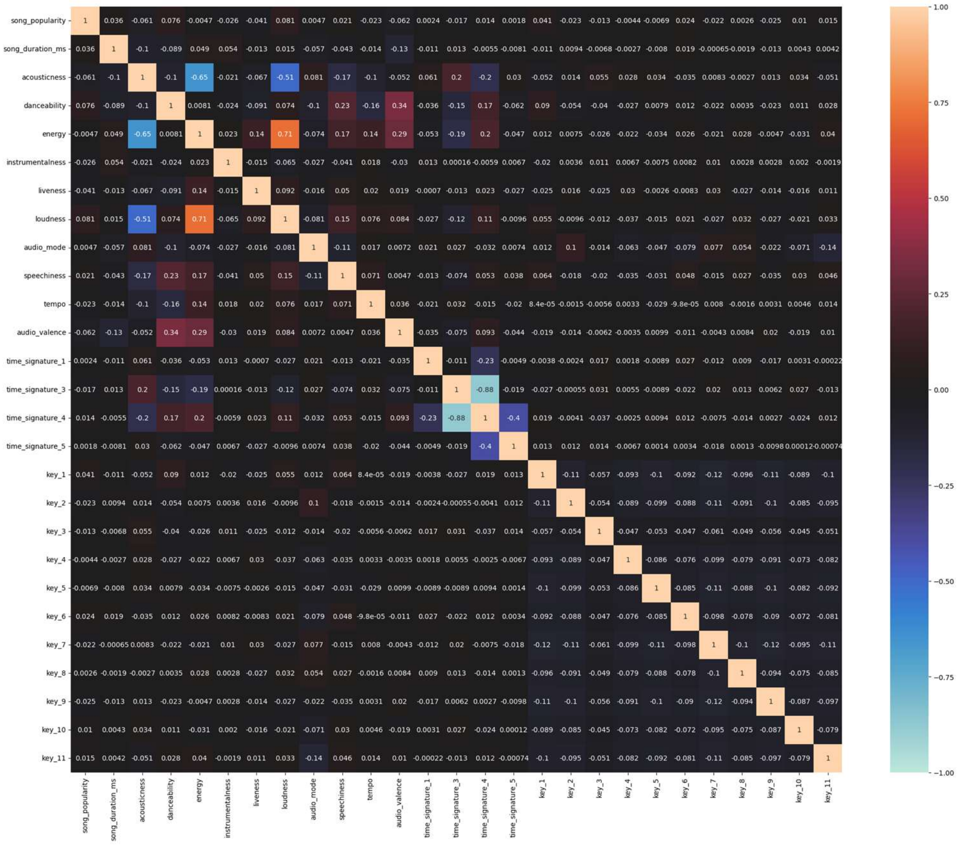

Figure 7: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

Inference: There seems to be strong multi-correlation between the features. Let us try to fix these...

Approach:

We can fix this multicollinearity with two techniques:
1. Manual Method - Variance Inflation Factor (VIF)
2. Automatic Method - Recursive Feature Elimination (RFE)
3. Feature Elmination using PCA Decomposition

4a. Manual Method - VIF

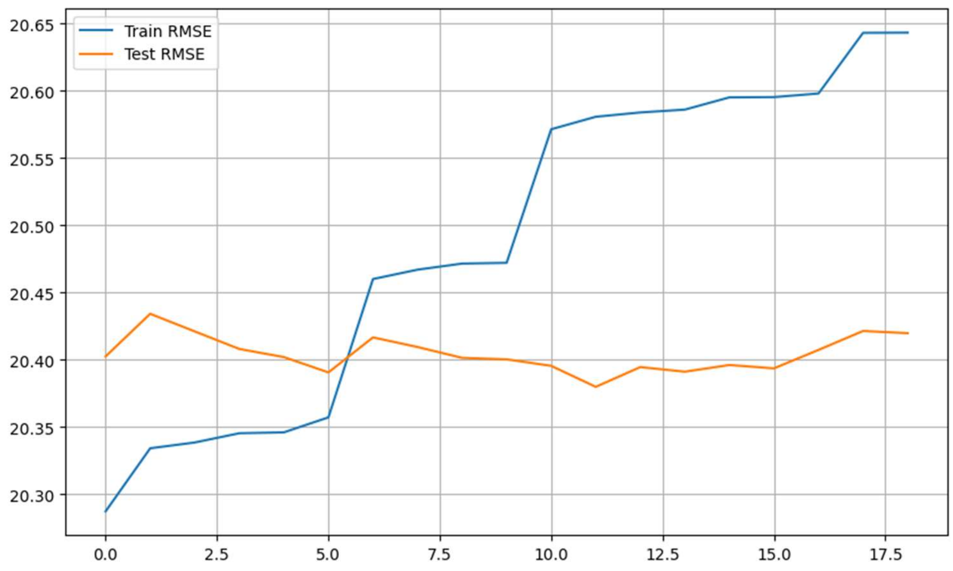

Figure 8: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

4b. Automatic Method - RFE

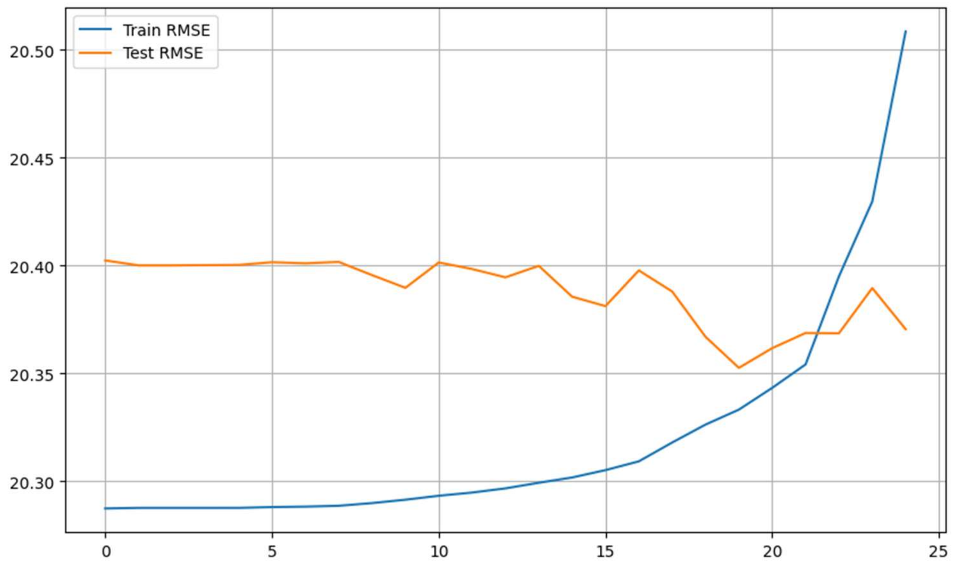

Figure 9: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

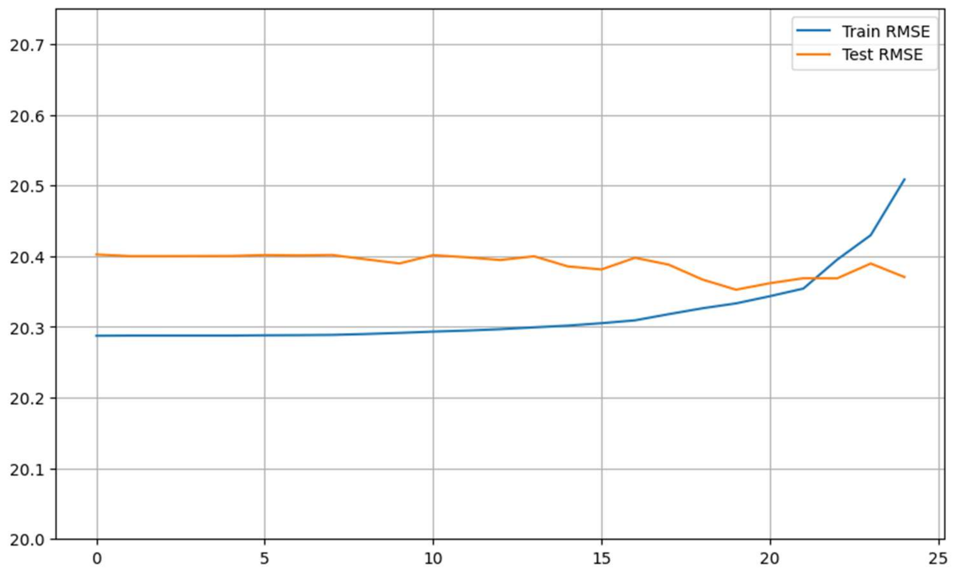

Figure 10: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

4c. Feature Elmination using PCA Decomposition

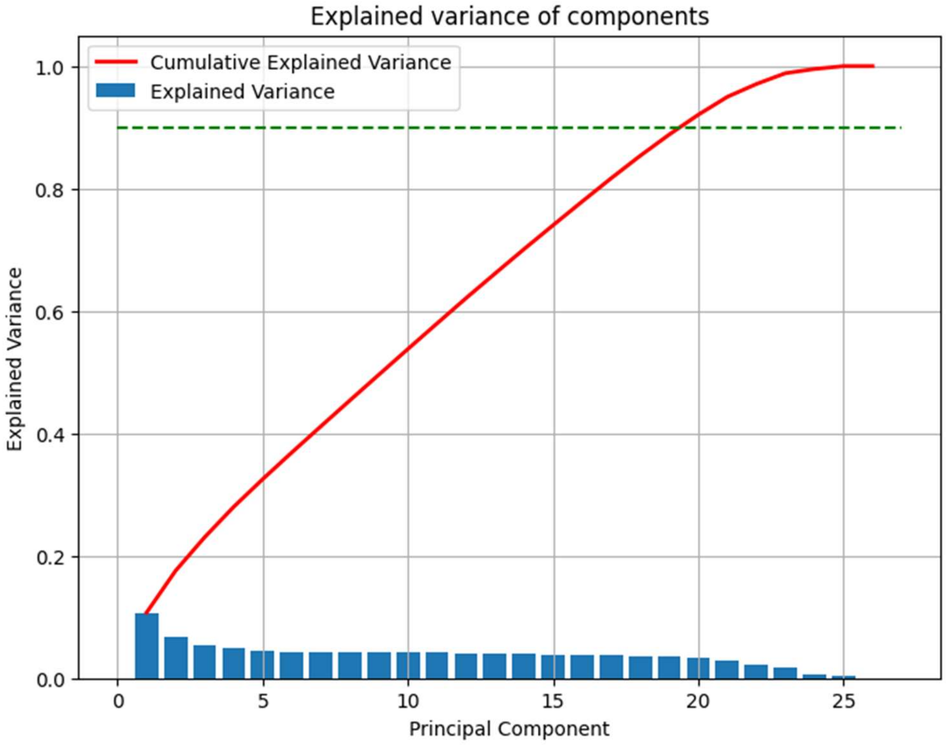

Figure 11: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

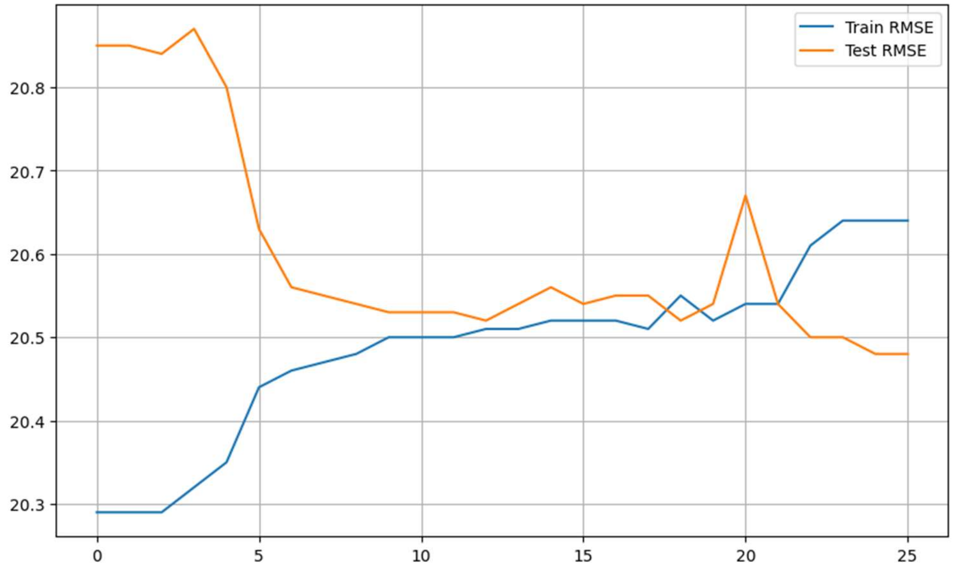

Figure 12: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

Inference:
It can be seen that the performance of the models is quiet comparable upon dropping features using VIF, RFE & PCA Techniques. Comparing the RMSE plots, the optimal values were found for dropping most features using manual RFE Technique. But let us skip these for now, as the advanced ML Algorithms take care of multicollinearity.

5. Predictive Modelling

Objective:
Let us now try building multiple regression models & compare their evaluation metrics to choose the best fit model both training and testing sets...

5a. Multiple Linear Regression (MLR)

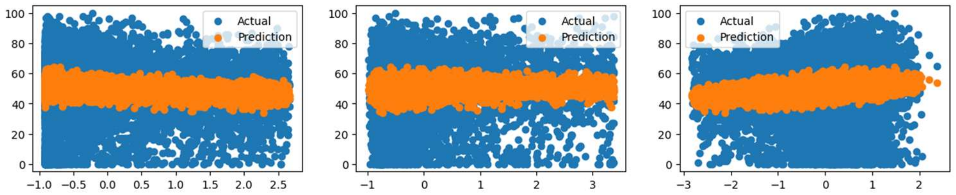

Figure 13: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

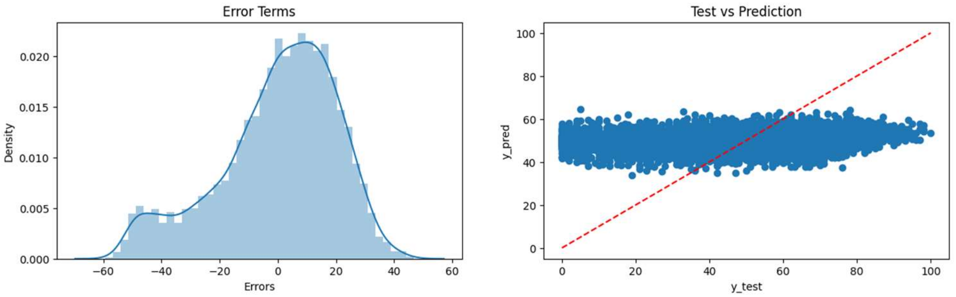

Figure 14: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

5b. Ridge Regression Model

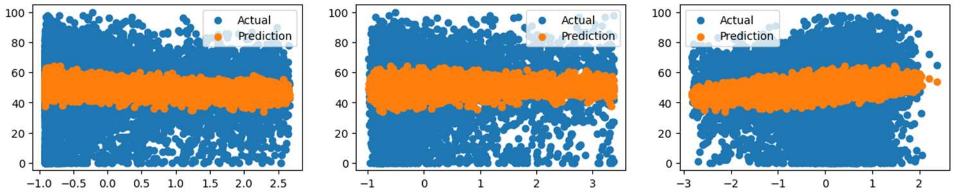

Figure 15: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

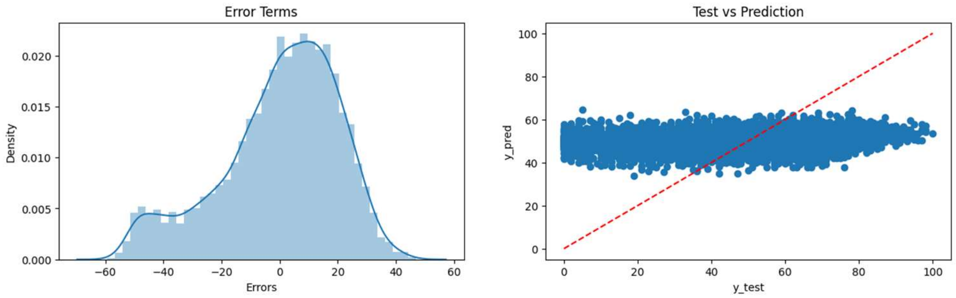

Figure 16: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

5c. Lasso Regression Model

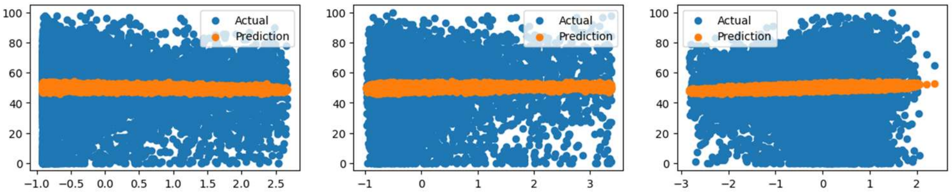

Figure 17: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

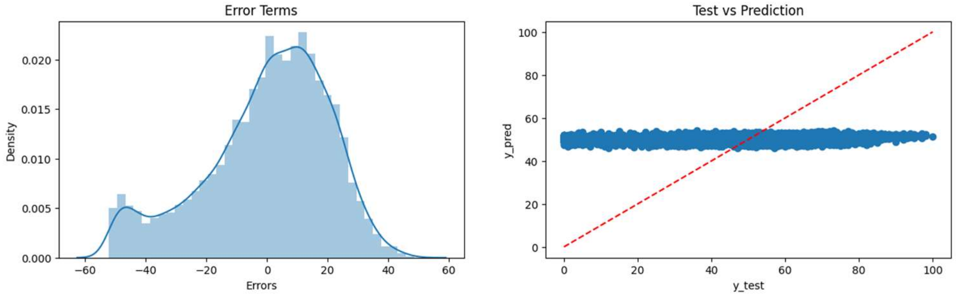

Figure 18: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

5d. Elastic-Net Regression

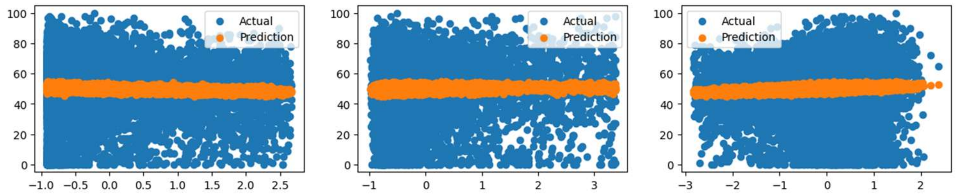

Figure 19: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

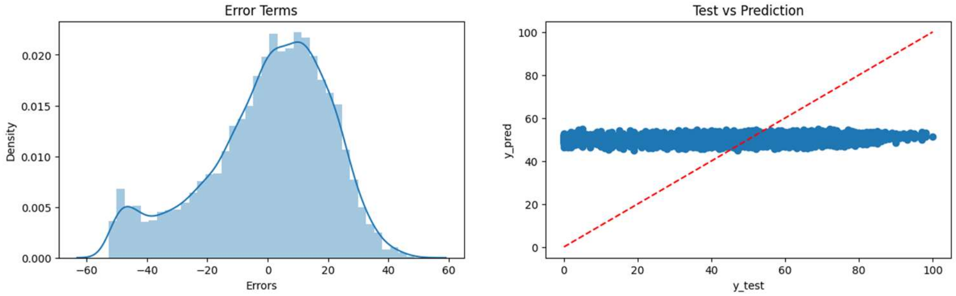

Figure 20: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

5e. Polynomial Regression Model

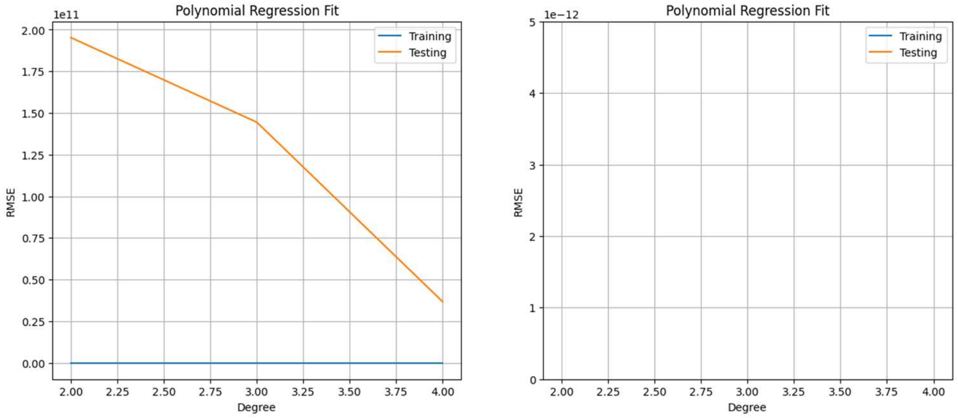

Figure 21: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

Inference: We can choose 2nd order polynomial regression as it gives the optimal training & testing scores...

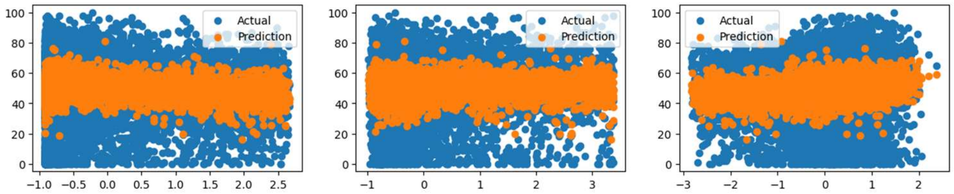

Figure 22: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

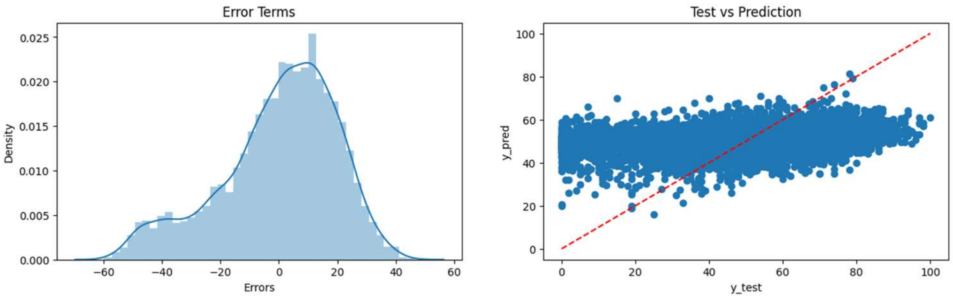

Figure 23: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

5f. Comparing the Evaluation Metics of the Models

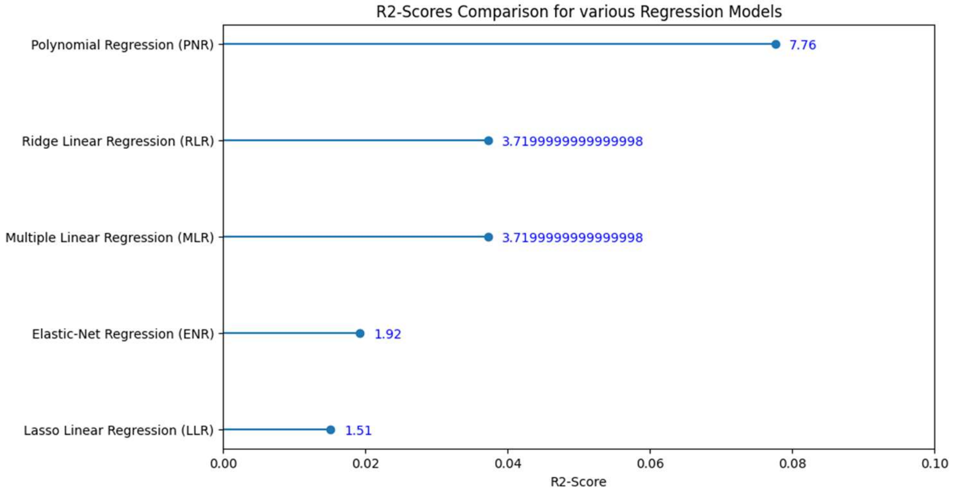

Figure 24: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

Inference: From the above plot, it is clear that the polynomial regresion models have the highest explainability power to understand the dataset.

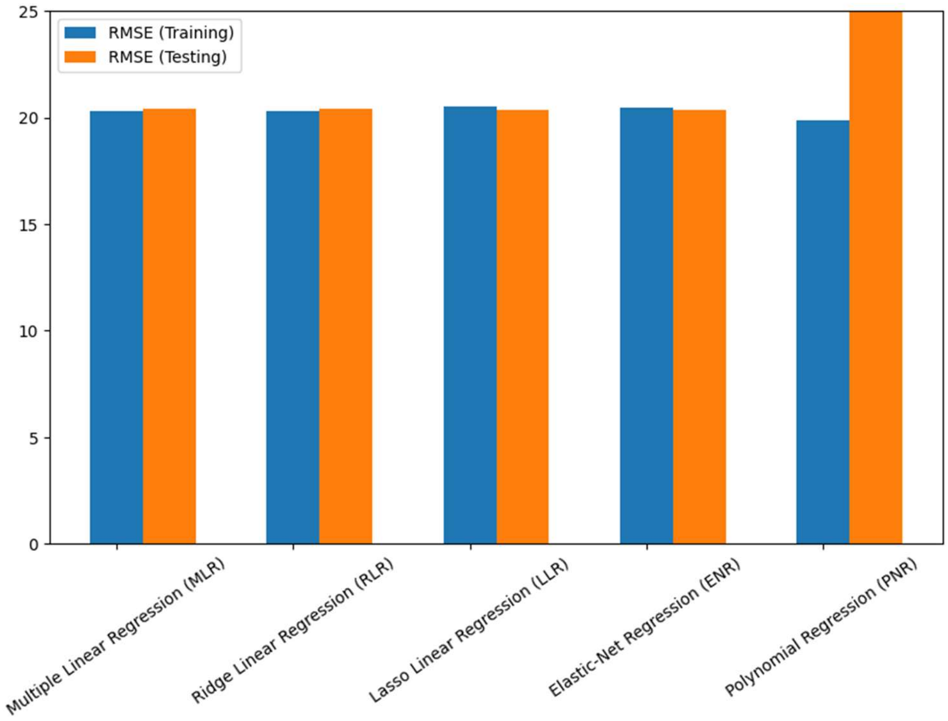

Figure 25: This graph represents model performance or data visualization. It showcases trends or patterns used to interpret prediction accuracy or feature importance.

The plot helps explain how certain variables affect the success or popularity of songs. Higher accuracy reflects better prediction capability.

Inference:

Lesser the RMSE, better the model! Also, provided the model should have close proximity with the training & testing scores. For this problem, it is can be said that polynomial regressions clearly overfitting the current problem. Surprisingly simple MLR Model gave the best results.

6. Project Outcomes & Conclusions

Here are some of the key outcomes of the project:

- The Dataset was quiet small with just 18835 samples & after preprocessing 33.4% of

the datasamples were dropped.
- Visualising the distribution of data & their relationships, helped us to get some insights on the feature-set.
- The features had high multicollinearity, hence in Feature Extraction step, we shortlisted the appropriate features with VIF Technique.
- Testing multiple algorithms with default hyperparamters gave us some understanding for various models performance on this specific dataset.
- While, Polynomial Regression (Order-2) was the best choise, yet it is safe to use multiple regression algorithm, as their scores were quiet comparable & also they're more generalisable.