Music Rating Prediction System

Final Project Report

Team Member	Team Member	Team Member	Team Member
Dhaval Sonawane	Jignesh Darji	Rohit Mandge	Saurabh sakpal
dsonawa@ncsu.edu	jndarji@ncsu.edu	rnmandge@ncsu.edu	ssakpal@ncsu.edu

1 Background and introduction

Music rating prediction system intends to predict how much rating a user would give on a scale of 1-100 for an EMI company music track they are made to hear. This helps the company know how a track or an artist would be received upon launch and helps them to design a music recommendation system by predicting songs for which the user will give a high rating.
This prediction must be made considering various attributes of users and artists. The data available in the data set about various users and artists will be merged into a single data frame. The resulting data frame will input to various regression models and their outputs will be analyzed. The model parameters for each of the models will be varied and the best suited values will be selected.
The resulting models will be validated against the validation data set. The performance of each of the model will be analyzed using the error rates captured from validation set. Among these models, we will choose the model that is best able to identify the interesting characteristics available in the data set such as correlation between the ratings and the words used, significance of various attributes, how attributes contributed to ratings, etc. The chosen model will be used to predict the values of the test data set and the error rates of the test data set will be captured and the performance of each of the models will be evaluated.

2 Proposed methods

2.1 Data

User data: User data in the users.csv file has information about user specific interests in music. It contains:

- Information about duration for which the users listened to purchased music and how long they listened to other music tracks.
- Demographic information such as age, gender, employment, etc.
- Answers to a set of 19 questions whose values are numeric in the range 0-100.

Artist data: Artist specific information is available in the words.csv file. It contains:

- Information about different users' description for the artist.
- Each row has list of 82 adjectives (0 indicating the adjective was used and 1 indicating the adjective was not used to describe the artist).
- "LIKE_ARTIST" attribute indicates users' rating for the artist on a scale of 0-100.

Training/Test data: Training/Test data from train.csv/test.csv has following information:

- Users' rating of the tracks of various artists.
- 4 attributes: User ID, Artist ID, Track ID and Rating on a scale of 0-100

2.2 Data Preprocessing

First challenge for data preprocessing is to merge the 3 data sets into a single data frame that consists of the user and artist information. We merge the datasets by using the user ID and artist ID.

Handling of missing values is done in the following way:

- For numerical attributes, we replace them with the median values.
- For categorical attributes, we replace them with the mode for the categorical attribute.

2.3 Methods

2.3.1 Linear Regression model

Linear regression is an approach for predicting a dependent variable y based on one or more explanatory variables (or independent variables) denoted X. In our problem statement, the ratings for a new track is the dependent variable and all the attributes describing the user and artist are explanatory variables.

Although the linear regression model can identify the correlation between ratings and some of the attributes, the model fails on many occasions. The model is tested against the validation data set and is found to have a high RMSE value of 27.4545. Hence we conclude that the linear regression model is not suitable for the given complex data set.

2.3.2 Linear model split by artist

As seen in the section 2.3.1, the linear regression model does not work very well for the given complex dataset. To solve this, we have attempted splitting the dataset by various artists and training a linear regression model for each artist. This significantly reduced the complexity of the data for the linear regression model.

After combining the results from all the models, RMSE was found to be high for this approach at 25.9347. This error rate is slightly better than linear regression on the entire dataset. However, this approach is not advisable as it does not account for any unknown artists in the testing set.

2.3.3 Gradient Boosting Model

Gradient boosting is a machine learning technique which produces a prediction model by combining weak prediction models, typically decision trees. Each successive tree is built for predicting residuals of the preceding tree. It builds the model in a stage-wise fashion, and generalizes them by allowing optimization of an arbitrary differentiable loss function. Shrinkage is used for shrinking the impact of each additional fitted stage. It reduces the size of incremental steps and thus penalizes the importance of each consecutive iteration.

For our problem statement, we built a gradient boosting model with shrinkage = 0.08 and interaction depth = 10 and maximum number of trees = 250. Lower shrinkage level means higher the time to train the model and higher the interaction depth means higher the depth of each of the trees. GBM model could predict the significance of each of the attributes in determining the prediction ratings of songs. Shrinkage value is kept as low as possible to avoid overfitting of the data.
GBM's error rate is lower than linear regression with an RMSE of 24.5728.

2.3.4 Random Forest

Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests improve over decision trees by avoiding overfitting

to their training set. Random forest models are known to be very good suit for complex data models.

The error rate of Random Forest is comparable with GBM with an RMSE of 24.7854

2.3.4 Random Forest split by artist

As we observed in 2.3.2, the results for linear regression improved as the data was split by artists. Since the data is complex, the resulting random forest model turns out to be complex. Splitting the data for random forest by artist reduces the complexity of the data and ultimately the model. The resulting error rate was found to be lower than applying Random Forest on the entire data.

Random Forest split by artist predicts the music rating with an RMSE value of 23.9862

3 Experiment and Results

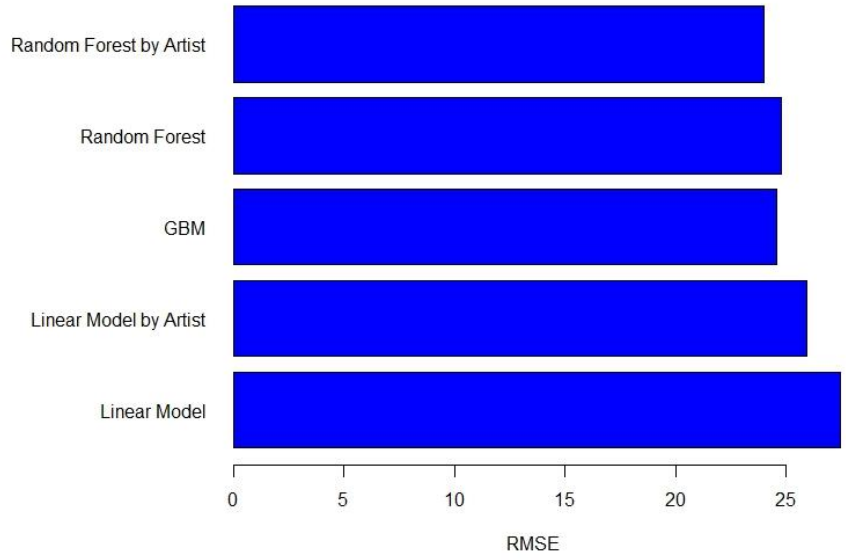
In the process of building the regression models, we could discover some important characteristics about the data.

3.1 Error rates

RMSE values for prediction of various models against the test data:

Model	RMSE Value
Linear Model	27.4545
Linear Model by Artist	25.9347
GBM	24.5728
Random Forest	24.7854
Random Forest by Artist	23.9862

Below is the bar plot for the same:



3.2 Artist Based Prediction

We tried to group data based on artist and build separate models for each group. This improved our accuracy in Linear Models and it will improve the accuracy in random forest too. The RMSE values improved by a significant amount. As we can see in the table above, artists based predictions are proved to be useful.

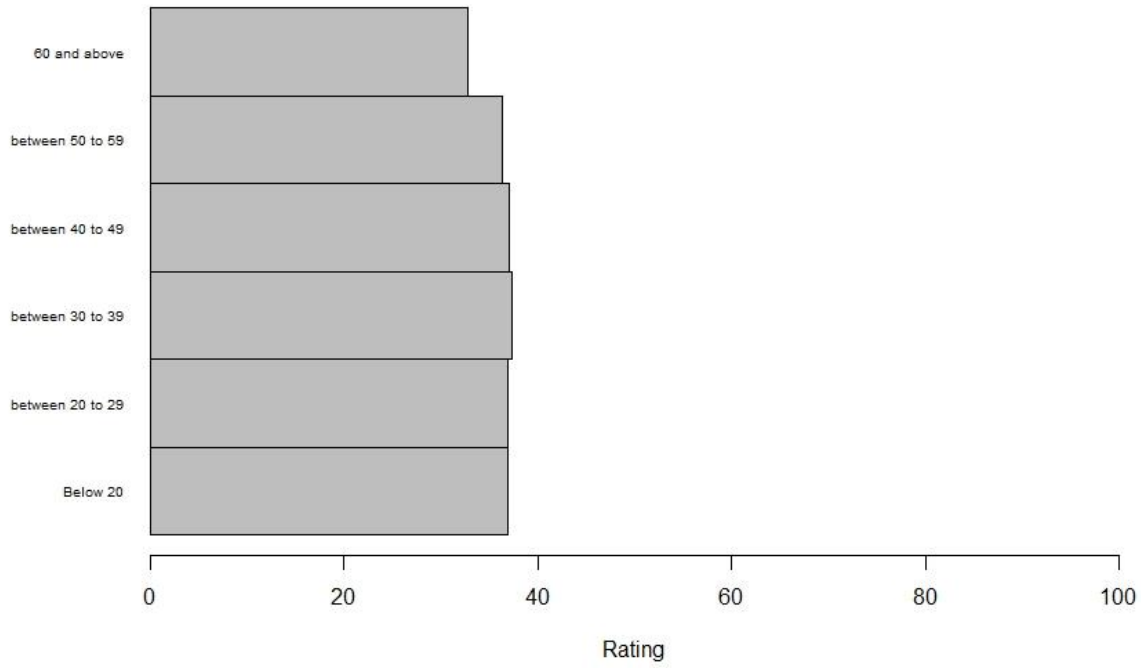
3.3 Track Ratings

Demographic Information based prediction: We tried to use the demographic information about the users and try to come up with some predictions as shown below.

3.3.1 Track Ratings by Age of User

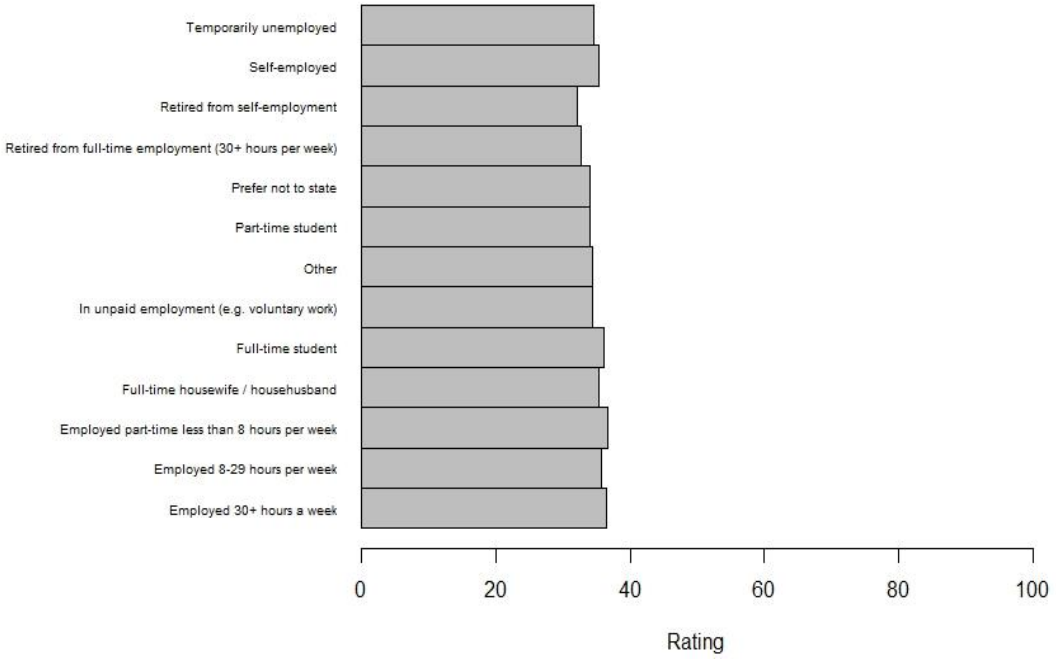
We are trying to divide the age group of users into 6 buckets and plot the average ratings provided by these age groups. Based on these average ratings we will try to predict ratings of new tracks. Below is the bar plot of average ratings by users as per their age groups.

The total average rating of all the age groups is **36.42975**.



3.3.2 Track Ratings by User Work Information

Here the data was grouped by User work information and the average ratings will be calculated for each group. Based on these average ratings, new track ratings were predicted. Below is the bar plot of average ratings by users as per their occupation.



3.3.3 Track ratings by user similarity

We found similar users with the help of Pearson correlation between every pair of users. Values for all pairs of users above a certain threshold will be considered to be similar. To predict the rating of a song by a user we calculated the mean of ratings given by all the similar users to the given user. This method did not work as there were not enough rating for a track by many users.

3.4 Visualizations

We tried to find the important adjectives used for the tracks based on the user ratings and gender. Below are the word clouds visualizing these adjectives.

3.4.1 Gender based

Word Cloud for Males vs Females respondents with respect to the kind of words they use. As seen, the top three words for Males and Females are the same: **Distinctive, Catchy and Good.Lyrics.**

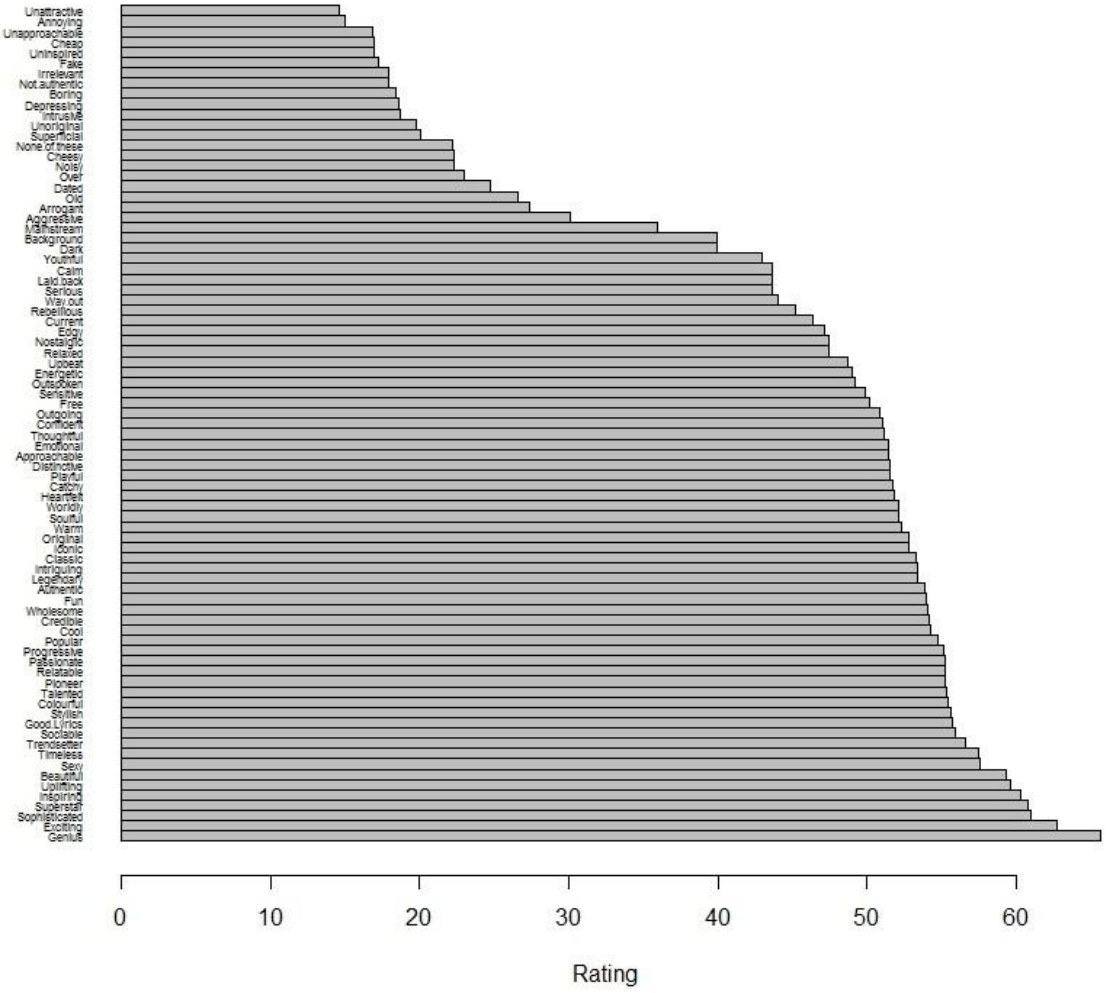
For Males

For Females

3.4.2 Track Rating based

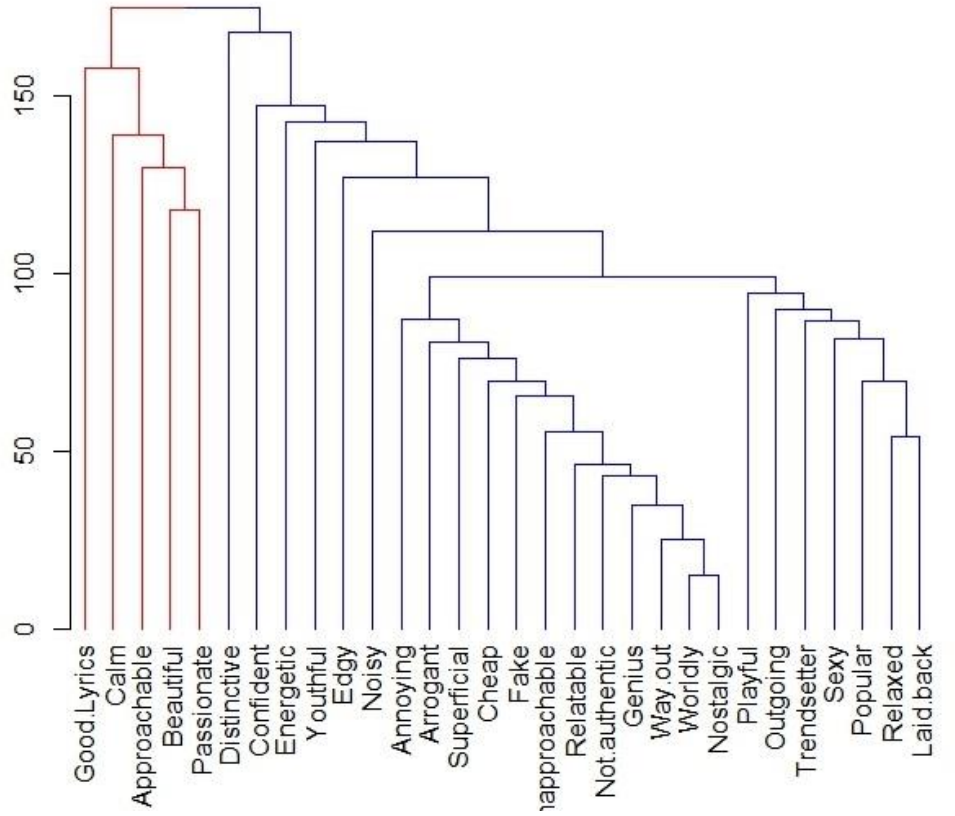
Word Cloud for important adjectives based on the track ratings given by users. As seen, the top 5 words are: **Genius, Exciting, Sophisticated, Superstar and Inspiring**.

The below bar plot shows the average ratings for songs when a particular adjective is used to describe it.



3.4.3 Word Similarity Hierarchy by Artist Word Co-Occurrence

Below dendrogram shows the similarity/correlation between adjectives used by the users to describe the tracks by an artist. It helps more in predicting the rating for a given track based on the user's description.



4 Conclusion

We have tried three different techniques to predict the track ratings: Linear regression, Gradient boosting and Random Forest. Because of the complexity of the data, linear regression does not perform very well. Gradient Boosting performs better than linear regression. However, random forest yields best results as it is well suited for complex data sets. Also, the results for linear regression and random forest can be improved by splitting the data based on artist. As we can see, the RMSE values was improved significantly when we split the dataset by artist.

5 References

[1] Majumdar, Abhishek, Arvind Kumar, and Sriram Manohar. "Music Recommendations Based on Implicit Feedback and Social Circles: The Last FM Data Set." N.p., n.d. Web

[2] Schneider, Astrid, Gerhard Hommel, and Maria Blettner. "Linear Regression Analysis." N.p., 5 Nov. 2010. Web

[3] Natekin, Alexey, and Alois Knoll. "Gradient Boosting Machines." N.p., 4 Dec. 2013. Web

[4] GitHub Code URL: https://github.com/mandgerohit/music-rating-predictor