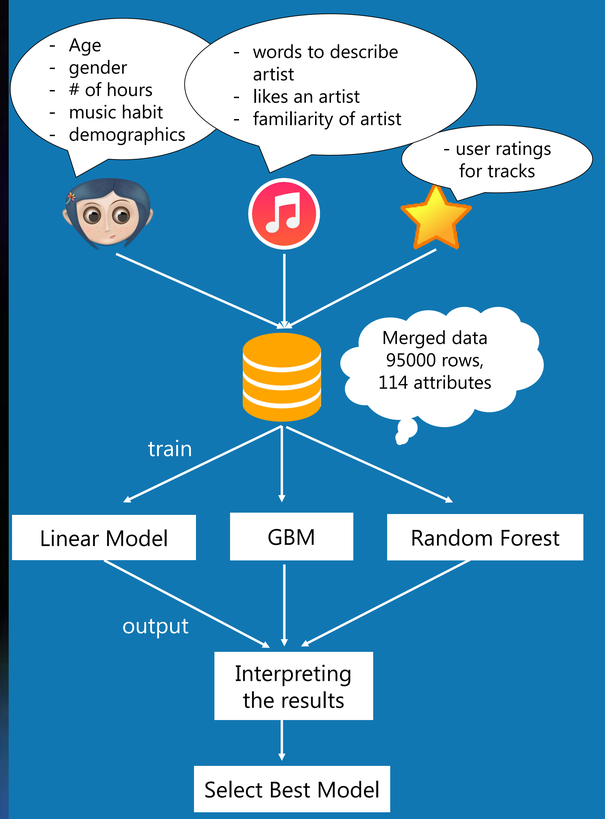
RATE MY MUSIC

Deepak Nandihalli, Gurudatt Palankar, Nitin Nagaraja, Soham Joshi

1. What this is about?

- We try to predict how much a user will rate a music track on a scale of 0 – 100.
- This could help in designing a music recommendation system by identifying tracks for which the user will give a high rating.

2. What we did?



3. What we tried?

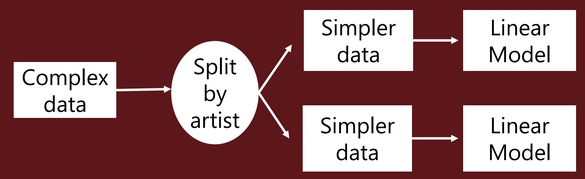
3.1. Linear Model

- Successfully identified correlation of few attributes with the ratings
- Correlation coefficient values of some attributes from the linear model:

Words	Correlation coefficient	
Catchy	5.21	✓
Not authentic	1.85	✗
Noisy	-5.353	✓
Legendary	-1.187	✗

- Turned out to be too simple for the complex data set.

3.2. Linear Model split by artist



- Compared to the linear model, which gave an RMSE of 27.61, we saw an improvement with an RMSE value of 24.76 for this model.

3.3. Gradient Boosting Model

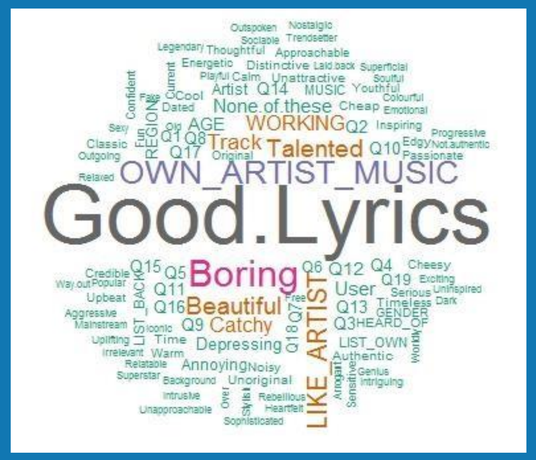
- Ensembles multiple decision tree models into a single model used to predict the ratings
- GBM is known to be good for handling perfectly correlated independent variables.
- The trained GBM model was able to correctly identify strong attributes that were significant in determining the ratings

Words	Significance
Good Lyrics	14.449
LIKE_ARTIST	9.58
Beautiful	8.876
Talented	8.3423
Catchy	7.7465

- The significance values indicate the extent to which they contribute in predicting the ratings for a new track

3.4. Random Forest

- Successfully captured the complex non linear functional dependencies of attributes.
- Building this model (with 114 attributes) was complex & computationally intensive.
- So, we pruned and selected the top 40 most significant attributes from the previous model's output.



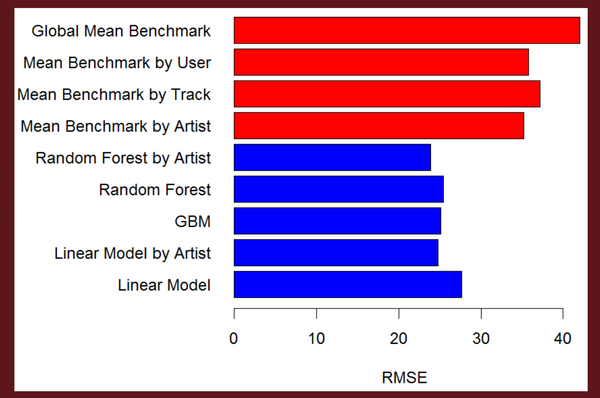
Word cloud visualization based on attribute importance

3.5. Random Forest split by artist

- We built a random forest model splitting by artist as we did with the linear model.
- As expected, we saw an improvement in the RMSE value which was 23.87, as opposed to 25.46 with the regular random forest earlier.

4. What we discovered?

MODEL ERROR RATES



We chose Random Forest by Artist as:

- Random forest seldom overfits the data
- Modelling by splitting on artist is able to capture the functional dependencies more intricately.

5. References

[1] Dataset: http://www.kaggle.com/c/MusicHackathon/data
[2] Gradient Boosting Model: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/
[3] Random Forest: http://www.math.usu.edu/adele/randomforests/uofu2013.pdf