Predictive Analysis Competition Presentation

✓ Critical Steps:

1. Data Exploration

- ✓ Adequate time spending on data exploration before starting modelling, causing a more structured data analysis process.

2. Data Tidying & Transformation

- Categorical Variables
  - ✓ All the categorical variables (including character and logical data types) have been transformed to dummy variables (numeric data type). It allows further machine learning models training.

3. Feature Selection: Reduced predictive modeling’s burden

- Variable Inter-set: Only keep common variables of all datasets
- Remove Near Zero Variance: Drop less predictive data
- Principal Components Analysis:
  - ✓ More efficient than subset selection or shrinkage
  - ✓ Generating linear combinations of original predictors, based on based on similarity measurement (eg. correlation)
  - ✓ Capturing a certain amount of variance

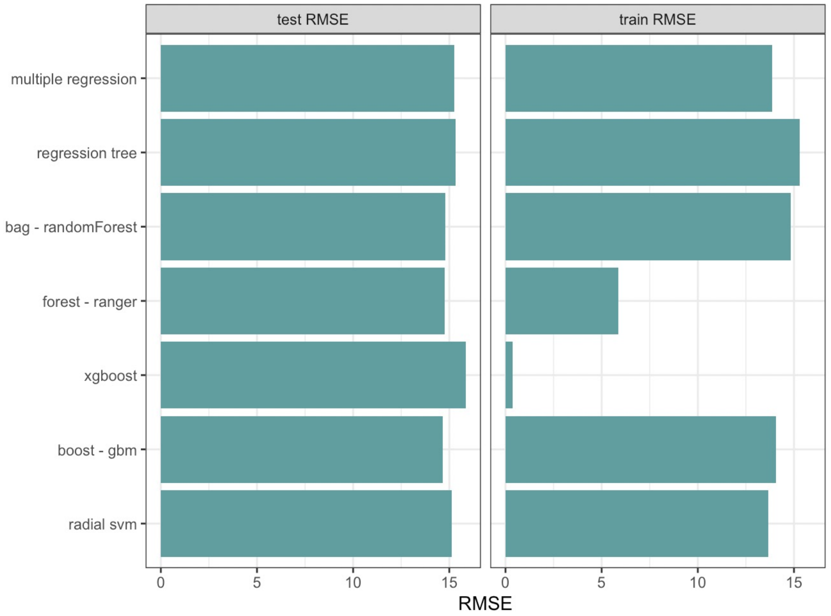
4. Data Analysis – Modeling

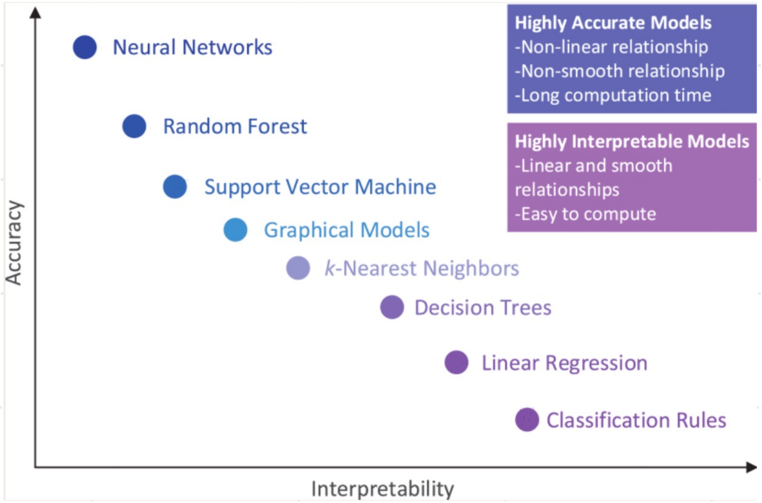
- ✓ I tried various models with default model parameters first. Each predictive model's RMSE (root-mean-square error) will be calculated to measure the model's accuracy for prediction. Then I compared these models' RMSE, and pick the models with lowest RMSE (best accuracy) for further parameters turning.
- ✓ Some models usually have higher flexibility and accuracy (eg. Bagging, Boosting, Random Forest, Support Vector Machine), while the other models have higher interpretability (eg. Linear Regression). Since our goal is to improve the predictive accuracy, the former are preferrable.

5. Data Analysis – Model Tuning

- ✓ Random Forest, Ranger, and gbm models are picked to further tuning. After an iterative process, the model’s prediction accuracy became better and better.
- Tuned Model Parameters
- Changed PCA Parameters
- Changed Seed

Model (without tuning) RMSE Comparison





Highly Accurate Models
-Non-linear relationship
-Non-smooth relationship
-Long computation time

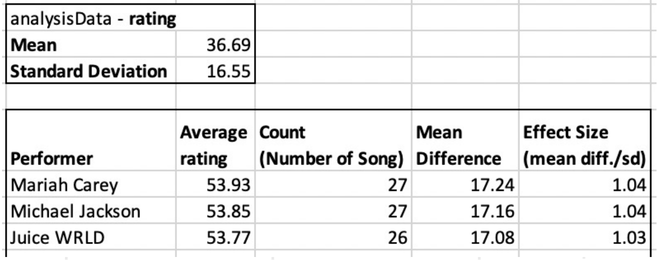
Highly Interpretable Models
-Linear and smooth relationships
-Easy to compute

× Mistake

- Feature Selection (Principal Components Analysis): I did a mistake in coding. I didn't include brackets while indicating the reduced number of components (1664*0.7) for PCA. Although PCA generates linear combinations of original 1664 predictors for predictive modelling, the number of predictors have not been reduced. Interestingly, I re-do the it with correct codes, the results are the same as before.

❑ Future Improvements

- Understanding the business indications of data set: Published literature and domain knowledge should help us a lot to understand what features are usually related or not related to a song’s rating.
- Feature Selection Approach
  - Correlation Check: I generated a correlation matrix for all predictors (1664) before PCA, but I gave up because it’s too much to check. However, I should at least check it before I created dummy variables.
  - Selection for the large amount of variables (eg. genre, perfomer), based on the certain variable’s effect to rating and its frequency of being a song’s feature. For example, filtering out performers with less than 25 songs in dataset, Mariah Carrey has the highest average rating. We should include “Mariah Carrey” as our predictors.



analysisData - **rating**	
Mean	36.69
Standard Deviation	16.55

Performer	**Average rating**	**Count (Number of Song)**	**Mean Difference**	**Effect Size (mean diff./sd)**
Mariah Carey	53.93	27	17.24	1.04
Michael Jackson	53.85	27	17.16	1.04
Juice WRLD	53.77	26	17.08	1.03

  - Model Tuning: If time and computer allow, I would like to enlarge the range of parameters for model tuning.