CP468 Project

Team Members: Stephen Morris, Lily Dinh

Intro/Main idea

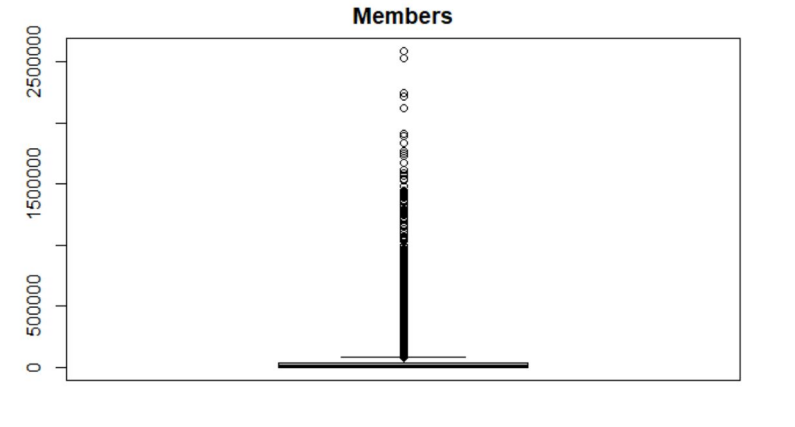
- The goal of this project is to find out if scoring affects an anime's ranking using a combination of R and Python.
- Anime recommendations in 2020 from about 17,000 anime collected from 300,000 users from MyAnimeList,
- R will be used primarily for exploratory data analysis while the model training will be done in Python.

Data overview:

- For our research we are primarily concerned with four variables, Score, Ranked, Members and Popularity. All other columns are dropped.
- The scaling of member counts are much higher than the other 3 columns, it is in the 10 thousands, so when using members count, we will modify it so that the scaling of members column will fit better with the other 3 columns.
- Data was cleaned and all NAN values were dropped. After the data was cleaned we were left with 11000 rows and 4 columns.

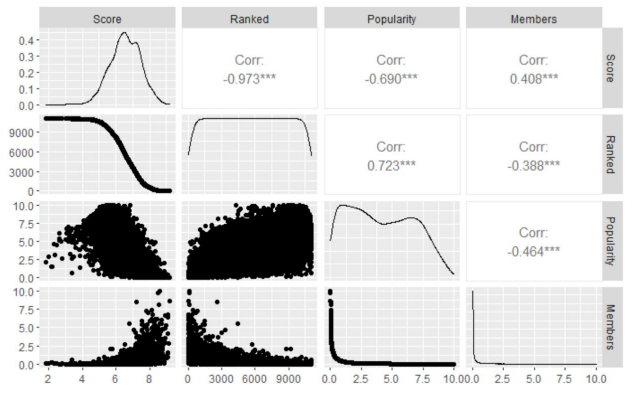
EDA findings:

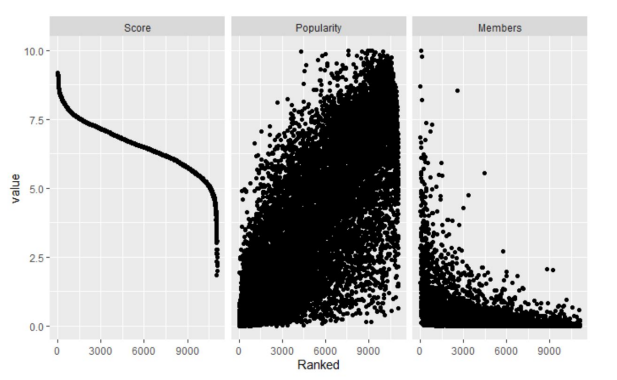
- Member counts have weak linear relationship with Ranking and many outliers
- Outliers can mislead the model, create inaccurate results



EDA findings cont.

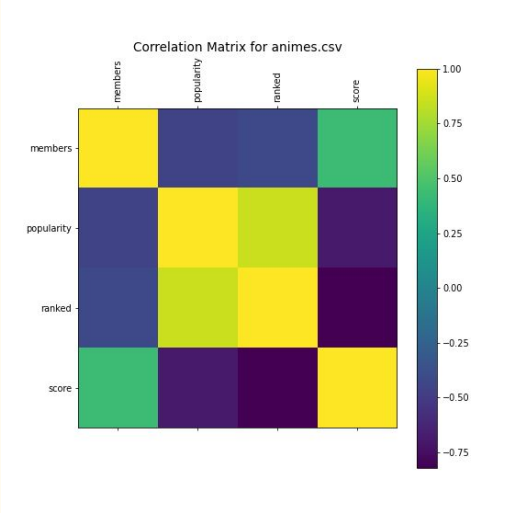
- Popularity has high variability, usually cause the model to overfit.
- Scores have strong negative linear correlation and only a few outliers

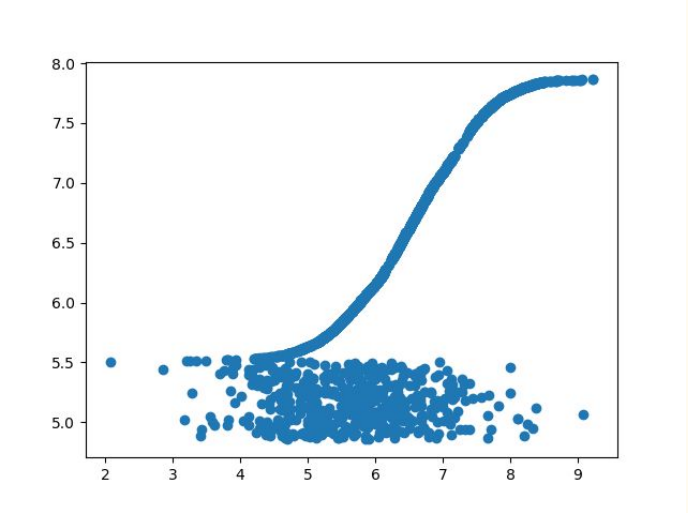




Training/Testing

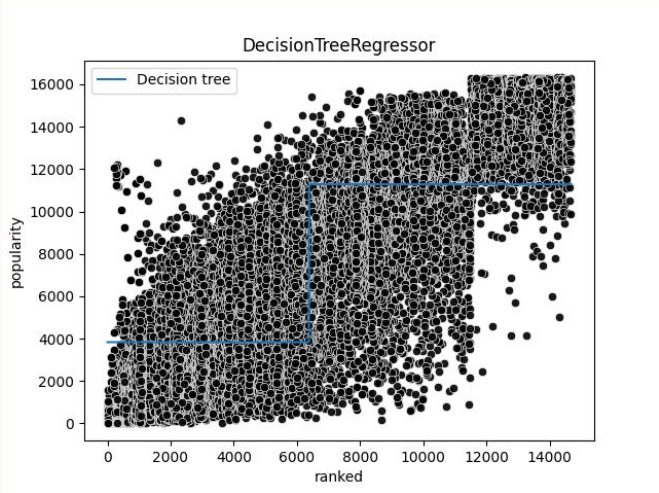
- Linear Regression Model
  - With linear model we achieved training accuracy of 71.25% and predictive data accuracy of 71.21%.
  - Absolute mean error = 0.360408168822885 23
  - Explained Variance Score = 0.7121042845447493

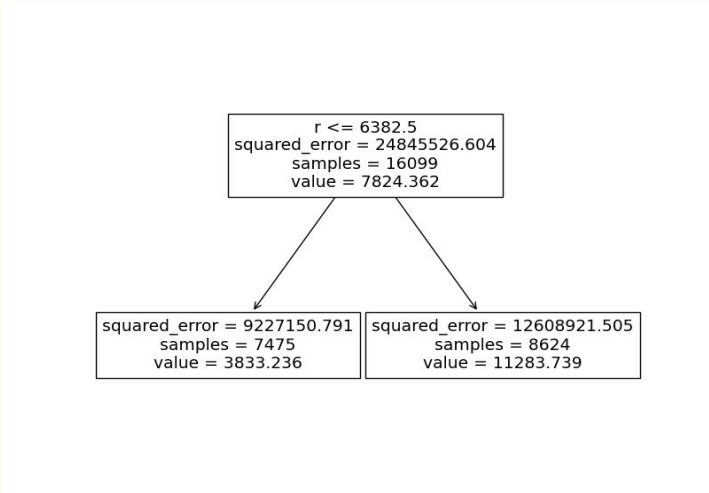




Training/Testing cont.

- Decision Tree model observations:
  - Decision Tree doesn't have a straight line to regress rank and popularity
  - No priori distribution
  - Feature space split into two partitions.





Logical Partition

Stephen Morris - Training/Testing

Lily Dinh - Exploratory Data Analysis