Abstract

As the global popularity of anime continues to rise, accurately predicting anime ratings has become essential for optimizing recommendations. Given data with various attributes such as categories, types, episode numbers, and the number of members involved, we achieved a Mean Squared Error (MSE) of 0.41 in our predictions utilizing a Gradient Boosting Regressor.

Data Preprocessing

Below are the first five rows of the data. Each entry provides information on the anime's name, genre, type, number of episodes, number of members, and rating.

anime_id	name	genre	type	episodes	rating	members
32281	Kimi no Na wa.	Drama, Romance, School, Supernatural	Movie	1	9.37	200630
5114	Fullmetal Alchemist: Brotherhood	Action, Adventure, Drama, Fantasy, Magic, Mili...	TV	64	9.26	793665
28977	Gintama°	Action, Comedy, Historical, Parody, Samurai, S...	TV	51	9.25	114262
9253	Steins;Gate	Sci-Fi, Thriller	TV	24	9.17	673572
9969	Gintama'	Action, Comedy, Historical, Parody, Samurai, S...	TV	51	9.16	151266

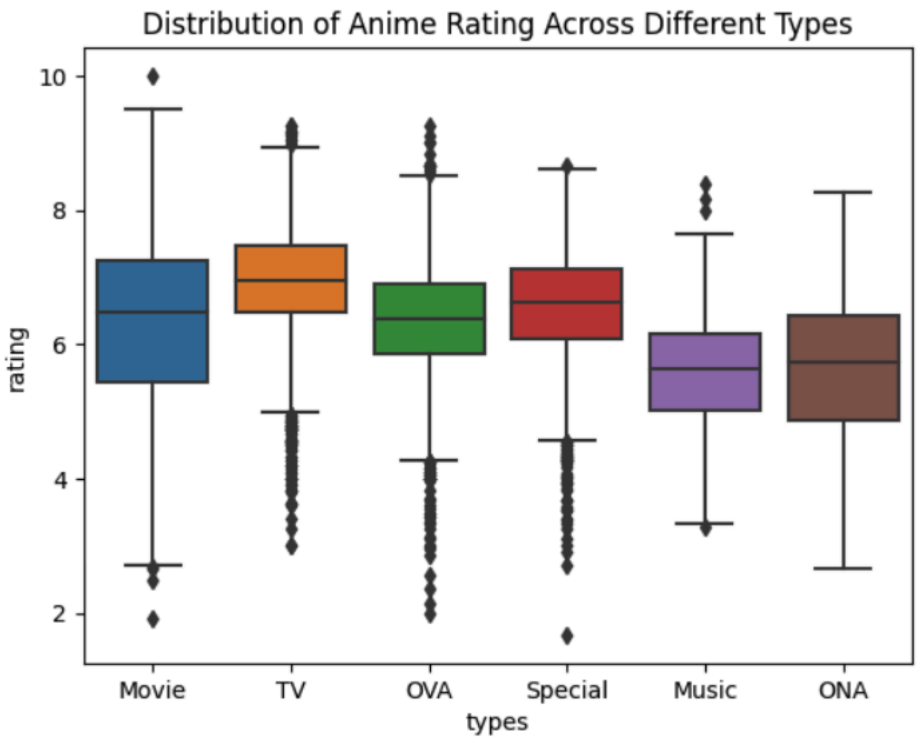
Data Wrangling

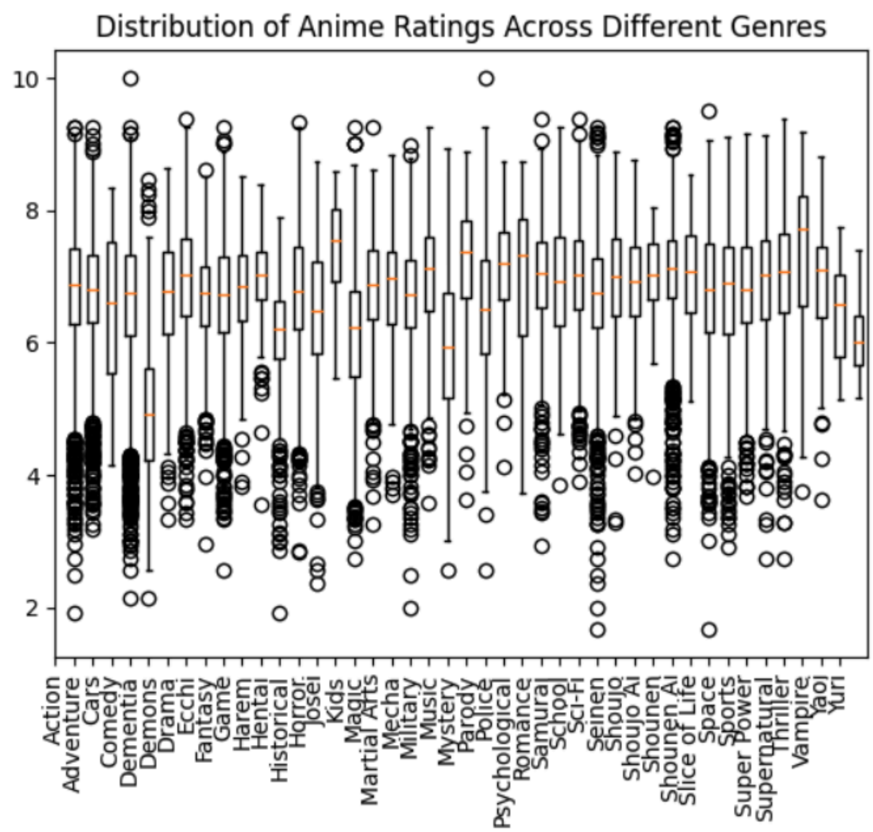
First, we checked for any missing values. While we had some missing values in the type, episode, and rating columns, we attempted to fill in the genre using information from other anime within the same series but with slightly different names, episodes, or types. Anime with `Movie` and `Music` types have a similar number of episodes, so we replaced the null episode values with the mean values for those anime. Finally, we dropped all the rows that did not have rating data.

We also ensured that each column type was as expected and that there were no abnormal feature values.

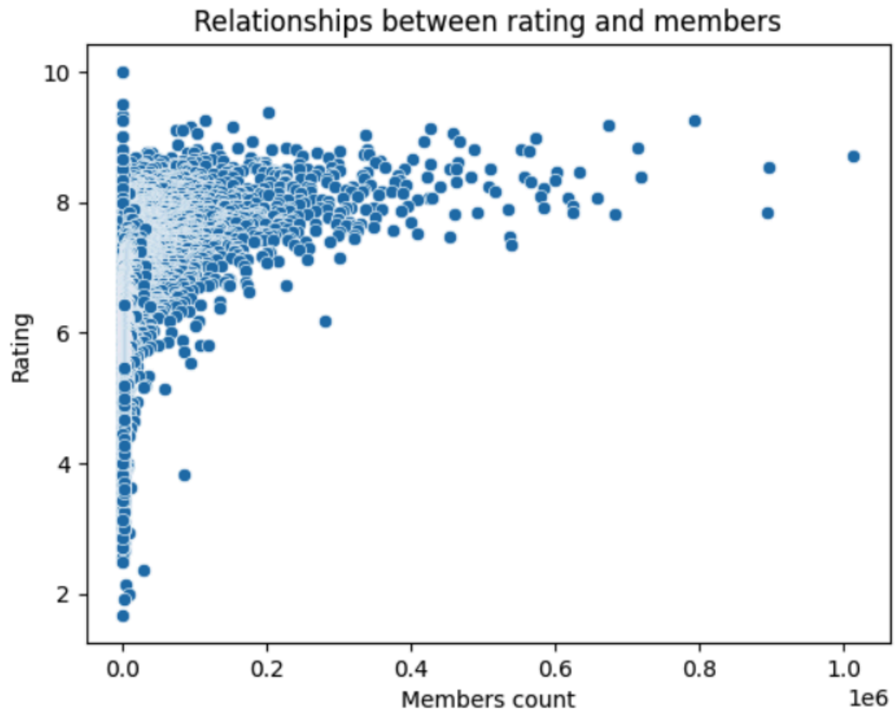
Data Exploratory Analysis

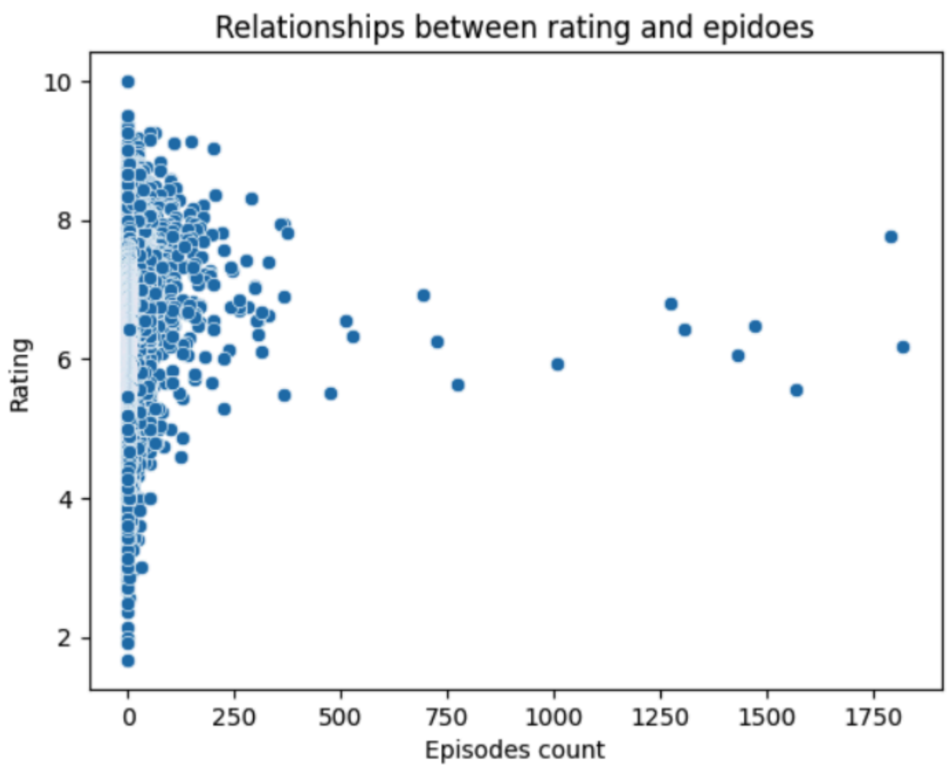
Based on the ANOVA results, the anime genre and type had F-values of 362 and 90, respectively, indicating that these two features have a significant impact on the rating.



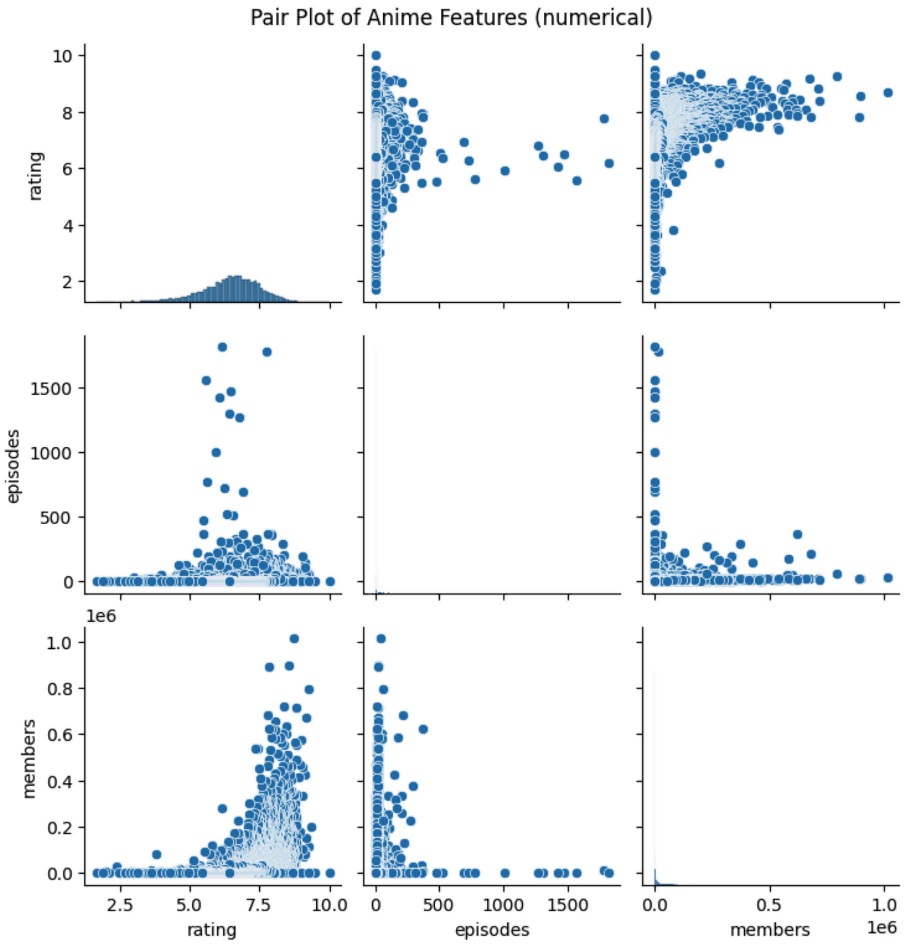


Anime episodes and members also show a slight positive correlation with ratings.



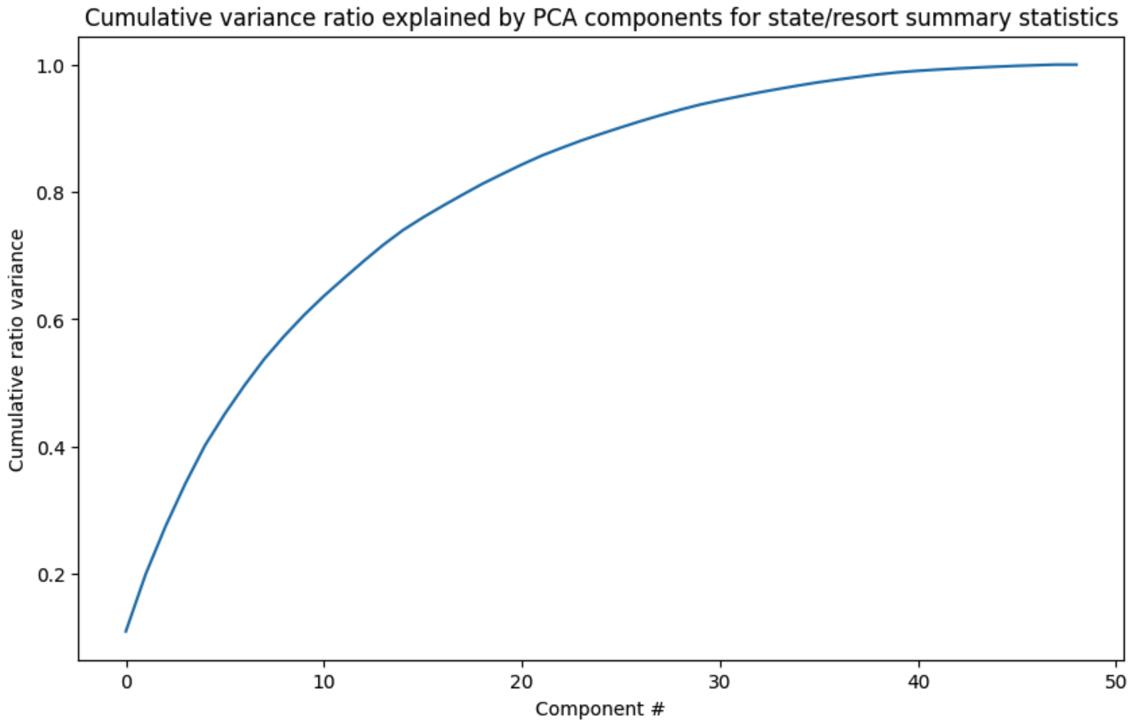


On the other hand, episodes and members didn't show a correlation with each other.



Pair Plot of Anime Features (numerical)
rating
episodes
members
1e6

The PCA analysis shows that we need a larger number of principal components to explain the large portion of the variance in the dataset.



Model selection & performance

To determine the best model for predicting ratings, we explored four different models:

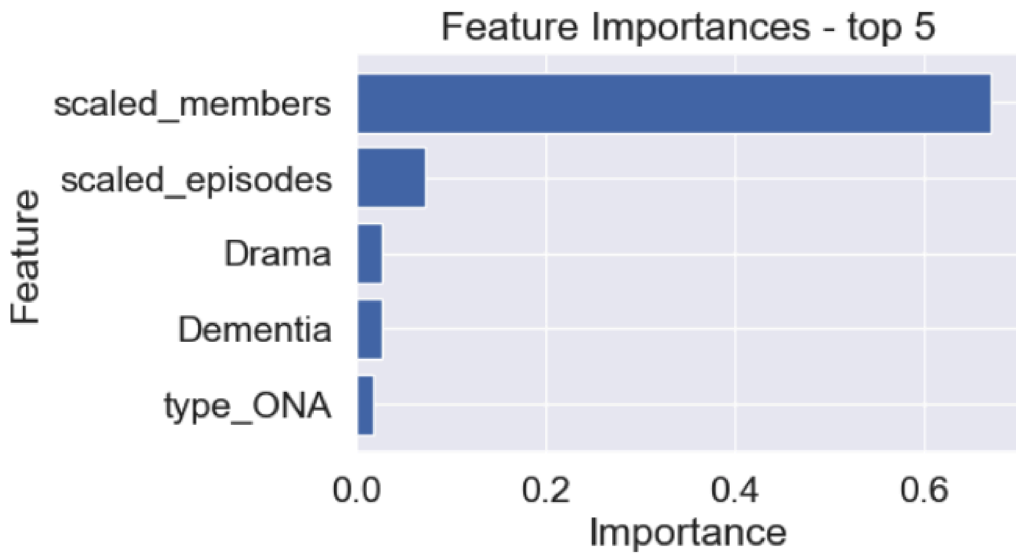
- Linear Regression
- Gradient Boosting Regressor
- Random Forest Regressor
- K Neighbors Regressor

Hyperparameters were optimized using RandomizedSearchCV.

Based on the results summary, Gradient Boosting Regressor emerged as the best model with the lowest MSE and highest R^2 scores.

	mse	r2
LinearRegression	0.661057	0.356994
RidgeCV	0.661492	0.356570
GradientBoostingRegressor	0.412962	0.598315
RandomForestRegressor	0.445549	0.566617
KNeighborsRegressor	0.537837	0.476849

Members and episodes are the two most important features.



Next Steps

Model Improvements

Consider experimenting with additional models, such as CatBoost and XGBoost, to potentially enhance prediction accuracy and performance.

Recommendation based on user’s rating history

There are limitations to predicting ratings based solely on the given features, as ratings can vary depending on personal preferences. Given that we also have data for each user's rating history, it would be beneficial to explore building a recommendation system specialized for each user.