Predicting Anime Ratings and Video Game Sales Using Machine Learning

Isac Lopez
CSCI 164 Final Project Summary

What's the Goal?

In this project, I explored how different supervised machine learning regression models perform on real-world entertainment data: anime and gaming. The goal was to predict **anime ratings** and **video game global sales** based on metadata, using classical machine learning approaches.

Datasets Explored:

Anime Dataset

- **Domain:** Entertainment
- **Description:** Metadata for anime (genre, type, episodes, member counts).
- **Objective:** Predict the average user **rating**.

Video Game Sales Dataset

- **Domain:** Gaming / Sales Analytics
- **Description:** Features like genre, platform, release year, critic scores.
- **Objective:** Predict **global sales** in millions.

Tools and Libraries Used:

- **Python**
- **scikit-learn** for ML algorithms and pipelines

- **pandas** & **numpy** for data handling
- **matplotlib** & **seaborn** for visualizations
- **GridSearchCV** for hyperparameter tuning

Machine Learning Models Applied:

- **Linear Regression** (baseline)
- **Random Forest Regressor** (tree ensemble method)
- **Multi-Layer Perceptron (MLP) Regressor** (simple neural net)

Key Results:

Anime Dataset:

Model	MAE	RMSE	R^2
Linear Regression	0.629	0.828	0.333
Random Forest	0.499	0.683	0.546
MLP Regressor	0.594	0.788	0.395

Interpretation:

- Random Forest achieved the best performance with the **lowest MAE (0.499)** and the **highest R^2 (0.546)**, showing it captured the relationships better than the other models.
- MLP Regressor also performed decently but required more computational time.

Gaming Dataset:

Model	MAE	RMSE	R^2
Linear Regression	0.592	2.037	0.012

Random Forest	0.590	2.042	0.008
MLP Regressor	0.589	2.042	0.008

Interpretation:

- All models struggled with this dataset, as shown by **very low R^2 scores (near zero)**.
- The models barely explained any variance, suggesting that simple features like genre and platform are **not strong predictors** of sales.
- Future work could involve richer features like marketing budgets, publisher size, or social media sentiment.

Comparison to Literature:

- **Anime Ratings:**
  Mohammed (2024) used hybrid collaborative filtering and deep learning to predict anime ratings with higher success. Compared to that, our classical models performed reasonably but could benefit from more complex techniques.
- **Video Game Sales:**
  Yang (2024) used ARIMA models for time-based sales predictions. Our static feature models did not perform well for predicting sales, indicating the **importance of temporal trends** in gaming markets.

Final Takeaways:

- **Random Forest** proved to be a solid, robust model on structured feature data (especially for anime).
- **Gaming sales** prediction is a **harder problem** without better or time-series features.
- Classical models like Linear Regression were good baselines but not sufficient for complex data relationships.
- Deeper learning approaches and richer feature engineering would likely be needed to significantly improve performance.

References:

Mohammed, Danishuddin (2024).
Predicting Anime Ratings and Popularity Using Machine Learning: A Data-Driven Approach.
ResearchGate.
https://www.researchgate.net/publication/386080233_Predicting_Anime_Ratings_and_Popularity_Using_Machine_Learning_A_Data-Driven_Approach_Collaborative_Filtering_Deep_Learning_and_Feature_Engineering_for_Accurate_Recommendations

Yang, Siying (2024).
Forecasts of Video Game Sales Based on the ARIMA Model.
EWA Direct Proceedings.
https://www.researchgate.net/publication/385728439_Forecasts_of_Video_Game_Sales_based_on_the_ARIMA_Model