DAMG 7370 Project Final Report

April 24th, 2024

Anime Data Analysis Group(Zeming Liu, Jiazhi Pang, Runchang Liu)

Team members:

The primary goal of this project is to ensure effective contribution from every team member. Each member takes responsibility for a distinct segment of the project as listed in the table:

NUID	Name	Role	Contribution Summary
002612666	Zeming Liu	Site Reliability Engineering	1. Project architecture design. 2. Deployment of all applications on Azure cloud and creation of pipeline. 3. Data preprocessing in databricks
002817310	Jiazhi Pang	Data Analysis and Exploration	1. preprocessing and exploratory analysis using PySpark 2. Built and tested predictive models. 3. Analyzed key features' impact.
002815600	Runchang Liu	Data Analyst	1. BI software implementation 2. Data analysis and production of visualizations 3. Dashboard creation

Role and Responsibilities

1. Dataset Introduction:

- Anime Dataset:

This dataset contains detailed information about anime series and movies, including titles, ratings, genres, synopsis, format (e.g., TV series or movie), episodes, airing status, production companies, animation studios, source material, duration, age rating, popularity, and user engagement metrics. It serves as a valuable resource for analyzing anime popularity, user preferences, and genre trends.

- Users Detail Dataset:

This dataset provides insights into anime viewers, including user ID, gender, join date, total days spent watching anime, average rating, and viewing habits such as series completed, on hold, dropped and planned to watch. It helps understand user-watching behavior and rating preferences.

- User Anime Score Dataset:

This dataset contains user-submitted ratings for various anime titles, capturing user interaction and preferences. It plays a crucial role in analyzing user engagement, trends, and overall anime reception within the community.

2. Objective:

Through the provided datasets, we can conduct various analyses to aid animation production companies and platforms in enhancing their operations and maximizing benefits. Here's a breakdown of sample analyses and more complex analyses:

Sample Analyses:

Top-Rated Anime: Identify the highest-rated titles to understand audience preferences.

Anime Genres Analysis: Explore the popularity and ratings of genres to guide content creation.

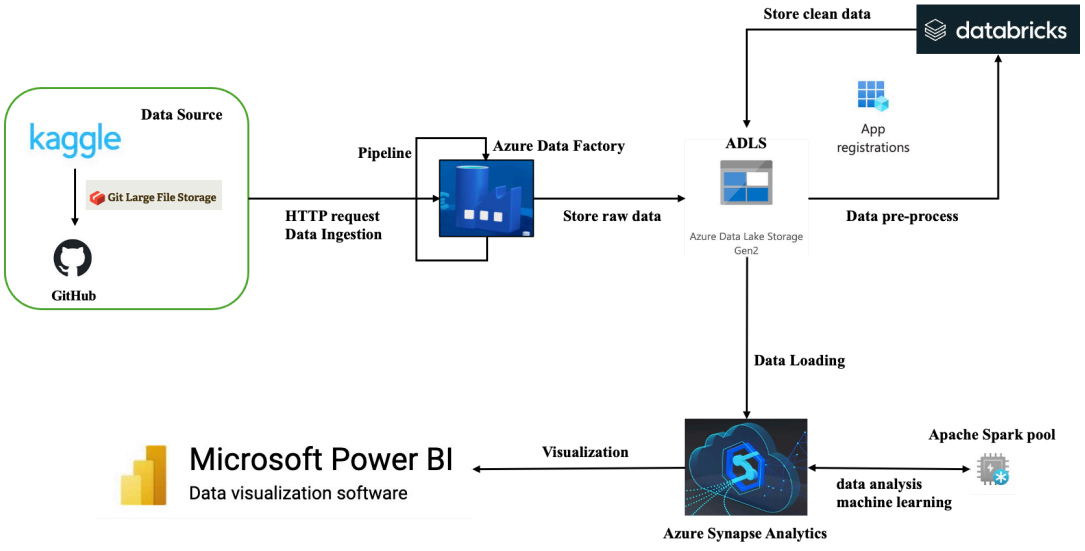
1. User Demographic: Analyze gender, age, and location data for tailored marketing.
2. User Watching Habits: Examine behavior patterns for optimized content delivery.
3. User Retention: Assess retention rates and engagement frequency for strategy refinement.

Deep Analyses:

1. Source Material Impact: Investigate how source material influences anime ratings.
2. Studios and Ratings: Analyze studio-user rating correlation for content investment.
3. Recommendation Systems: Develop ML-based systems for enhanced user experience.

These analyses empower companies to enhance operations and maximize benefits by making data-driven decisions.

3. Data Architecture:



Our project is deployed entirely on the Azure cloud, utilizing Azure Data Factory, Azure Databricks, Azure Data Lake Storage Gen2, Azure Synapse Analytics, and Power BI.

The whole workflow begins with downloading data from Kaggle. Due to the large size of the datasets, we use LFS, The Git Large File Storage, to upload the data to GitHub. Then, in Data Factory, we use HTTP requests to retrieve the data and store it in Azure Data Lake Storage (ADLS). Next, we preprocess the data using Databricks, perform analysis and secondary processing with Synapse Analytics, extract valuable datasets, and finally visualize the processed data using Power BI.

4. Pipeline

Use Data Factory to extract it by using HTTP requests. Since the data is a regular URL address, we use GET requests to retrieve it. Create a pipeline with three copy data activities corresponding to the extraction of three data groups. Connect to DataBricks and Synapse to run the notebook that handles the data.

5. Data Analysis and Exploration:

5.1 Data Preprocessing

We utilized the Apache Spark framework to ensure accurate parsing of data containing complex formats, such as multiline fields and special characters. Initially, we loaded the data from an Azure storage account into the Spark framework. Upon examination, we observed numerous parsing errors in the dataset.

In the preprocessing stage, we identified that by focusing on the 'episodes' column, we could effectively remove rows with irregular formatting, thereby addressing these parsing inconsistencies in the dataset. Subsequently, we performed data type conversions on key fields like 'episodes' and 'score' to ensure that the values in these fields were accurately parsed as floating-point numbers. Rows that could not be properly converted were removed. Similarly, we eliminated entries containing numbers from the 'producers' and 'studios' fields.

5.2 Data Exploration

In processing the 'genres' field, we separated each genre into its own row, as originally all genres for each anime were combined into a single column. facilitating further analysis by genre in the subsequent visualization stage. For instance, this arrangement allows us to easily count and compare the number of anime titles across different genres, thereby identifying which genres are more popular or associated with higher ratings.

5.3 Network Relationships and Connectivity Analysis

The network graph constructed reveals the strength of relationships between production companies and studios, and their associated anime works. For example, certain

production companies like "Toei Animation" and "Sunrise" exhibit a high degree of connectivity, indicating that they have produced a large number of anime works. This type of graph can help us identify key influencers within the industry and may indicate these entities' market control.

```
+--------------+------+
|        entity|degree|
+--------------+------+
|       unknown|  6606|
|toei animation|   576|
|       sunrise|   413|
|      tv tokyo|   376|
|        lantis|   332|
|production i.g|   291|
|     j.c.staff|   283|
|           nhk|   278|
| bandai visual|   278|
|      madhouse|   277|
+--------------+------+
only showing top 10 rows
```

Building on this foundation, by analyzing the weighted scores for each entity, we discovered that some lesser-known entities (such as "Miracle Robo") have exceptionally high average ratings. This phenomenon may indicate that niche or genre-specific studios are capable of producing high-quality works. These analytical findings provide valuable market insights for investors and content creators, guiding them to prioritize collaborations with these highly rated entities.

```
+--------------------+-----------------+
|              entity|        avg_score|
+--------------------+-----------------+
|        miracle robo|9.039999961853027|
|              quaras|  8.9399995803833|
|          voque ting|8.850000381469727|
|     studio signpost| 8.78000020980835|
|           seikaisha|8.550000190734863|
|   rex entertainment|8.539999961853027|
|               oniro|8.539999961853027|
|sony music solutions|8.490000089009603|
|              sumzap|8.479999542236328|
|                 fbc|8.470000267028809|
+--------------------+-----------------+
only showing top 10 rows
```

6. Model Training and Prediction Analysis

6.1 Feature Engineering and Data Transformation

By applying StringIndexer and VectorAssembler, we have transformed categorical variables such as producers, studios, type, and source into numerical indices and combined them with the member's column—which has been converted to an integer representing

audience size—to form feature vectors. This transformation enabled us to utilize these attributes in our machine-learning models.

6.2 Model Overview and Evaluation

We trained three different regression models: Random Forest, Gradient Boosting Trees (GBT), and Linear Regression, to predict the ratings of anime. We evaluated the predictive performance of each model using the Root Mean Squared Error (RMSE) on the test dataset.

The Random Forest model yielded an RMSE of 0.518562, indicating moderate predictive accuracy.

The GBT model demonstrated superior performance with an RMSE of 0.458022, suggesting its greater effectiveness in handling this dataset.

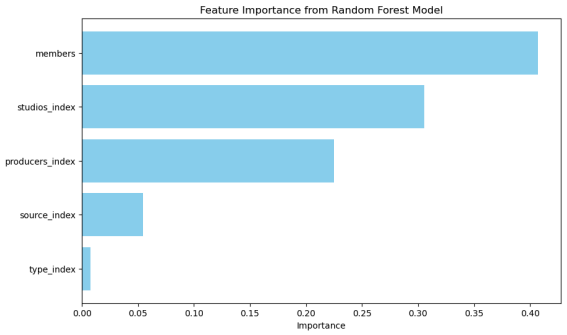
The Linear Regression model performed the worst, with an RMSE of 0.781675, possibly due to the data's nonlinear characteristics and complex interactions, which the linear model failed to capture.

6.3 Application of Ensemble Methods

Further, we experimented with an ensemble learning approach, averaging the predictions from the three models to form an ensemble prediction model. The ensemble model achieved an RMSE of 0.507526, which outperformed the standalone Random Forest and Linear Regression models and was close to the GBT model, illustrating the potential of ensemble methods to enhance prediction accuracy.

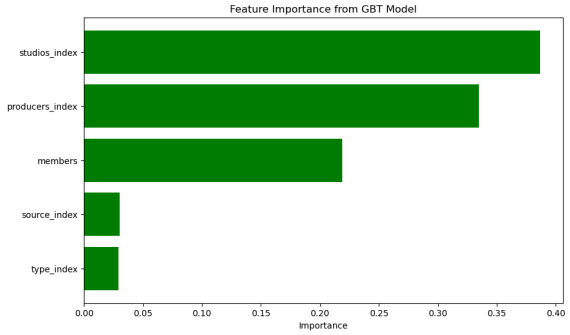
We also employed a bagging ensemble approach by creating five random forest models, each trained on a subset of the training data. The predictions from these models were aggregated, and the final ensemble prediction was obtained by averaging the individual predictions. This bagging ensemble model yielded an RMSE of 0.502446 on the test data, which is a further improvement from the single ensemble model. This result underscores the effectiveness of bagging in enhancing predictive performance, particularly in scenarios where the underlying data may have complex structures that can benefit from the diverse perspectives of multiple models.

6.4 Feature Importance

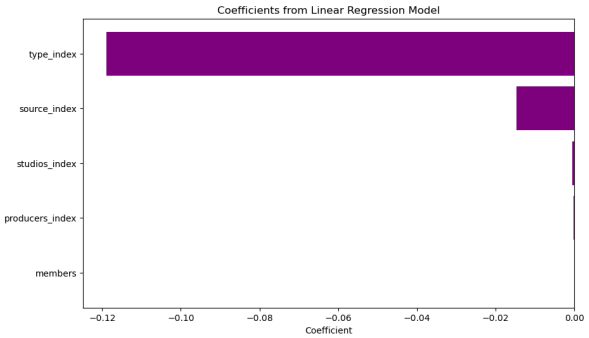


In the Random Forest model, the member's feature was of the highest importance, signifying that the number of viewers is a key factor affecting anime ratings. The popularity of an anime could directly impact its rating, as a larger audience size may encompass more positive reviews. Additionally, studios_index and producers_index also showed high

importance, hinting at the significant role of the reputation or quality of production companies and studios in ratings.



For the GBT model, studios_index and producers_index significantly rose in importance, becoming the most critical features for predicting anime ratings. This may reflect that the impact of production quality and studio reputation on ratings might be more complex, or there exists a stronger nonlinear relationship between them and the ratings. The GBT model, by accounting for these nonlinear relationships, was able to predict ratings more accurately.



In contrast, in the Linear Regression model, type_index and source_index had the highest absolute coefficients, indicating that the anime's type and source significantly affect rating variations. Notably, the coefficients from the Linear Regression were negative, which could indicate that higher index values in the used numerical encoding are associated with lower ratings, or these features have a negative correlation with the ratings.

6.5 Conclusion

Through detailed model assessments and feature analyses, we can better understand how various factors influence anime ratings and guide future data collection and model improvement strategies to optimize our predictive framework. These findings support continuous improvement and adjustment of model parameters to further enhance prediction accuracy.

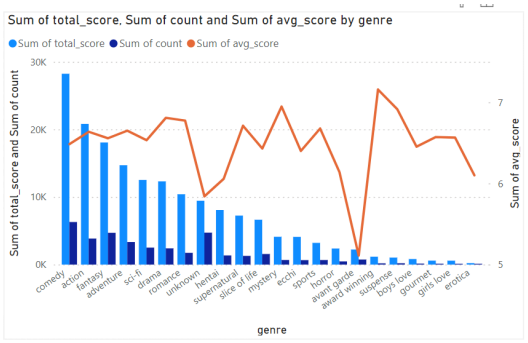
7. Visualization:

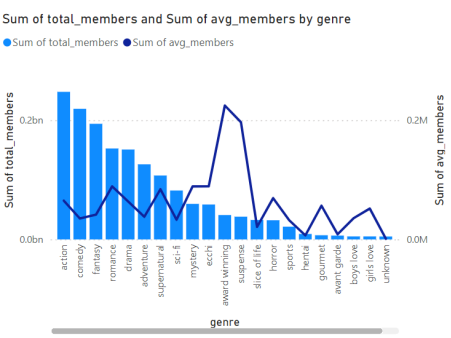
For the visualization part, we used the Power BI tool to visually present and analyze it through the analysis results of synapse in the previous step and created a dashboard for more visualization of the full range of data.

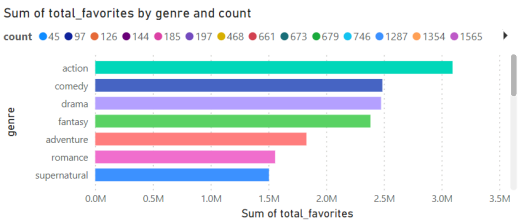
7.1 genre-based analysis

The bar chart depicts the clicked favorites for various anime genres. Comedy leads in total score and count, yet its average score doesn't stand out. "Boys love" and "girls love" genres show relatively high average scores. Surprisingly, The award-winning genre has the highest average scores, but there are very few of them, which may indicate that viewers are more likely to watch anime that have won awards, which tends to indicate that the anime has a high quality and reputation.

Regarding membership data, the action genre boasts the most members but not the highest average. Mystery peaks in average membership, indicating high engagement per work. "Slice of life" and "sports" genres show balanced membership despite lower total counts. The "Gourmet" genre has a low total number of members but a moderate average. The "Unknown" genre lacks clear membership data, possibly due to categorization issues or incomplete data. Understanding context and data collection methods is crucial for accurate analysis

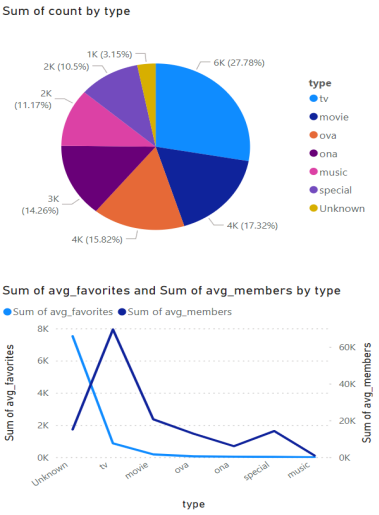






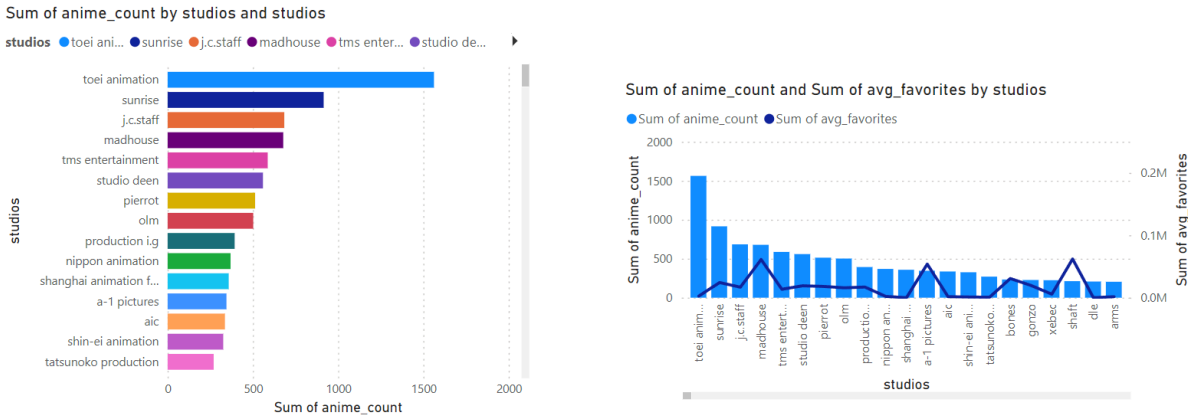
7.2 type-based analysis

Pie charts offer a straightforward view of anime distribution across various platforms. Comparing average favorites and membership for different genres, TV productions peak in average favorites but not in membership, suggesting a loyal fan base without broad appeal. The original video animation (OVA) genre peaks in membership but lags in average favorites compared to TV shows. Online animation (ONA) and specials show lower engagement overall. Music-related titles have high membership but low favorites.



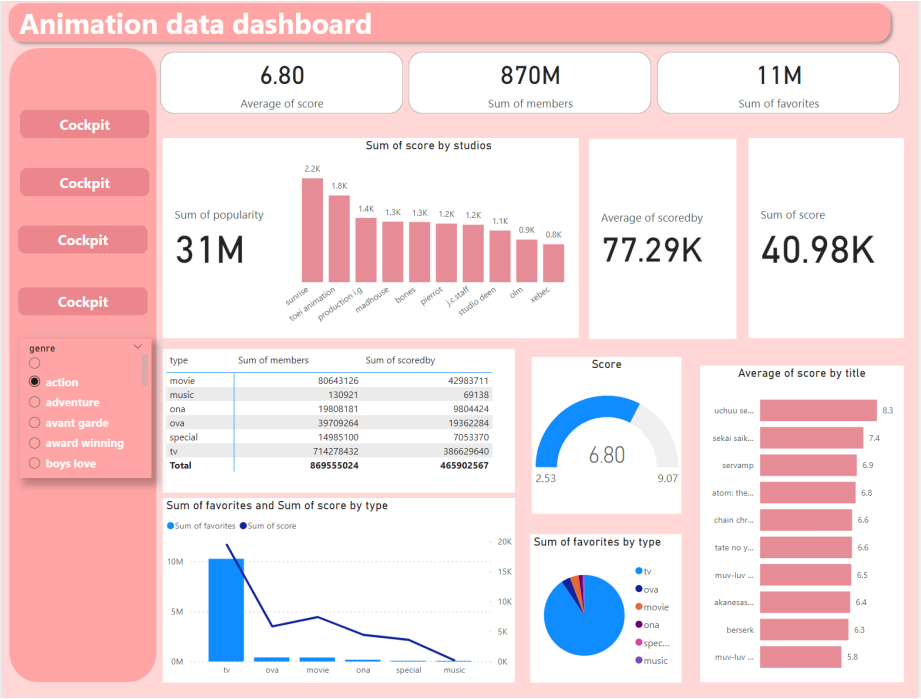
7.3 studio-based analysis

The right chart compares anime count and average favorites by production studios. The first studio leads in quantity but not in average favorites. Some studios with fewer titles have high average favorites, indicating popular releases. Others with many titles lack standout favorites, suggesting varying popularity. These insights help gauge studio performance, but factors like genre, marketing, and historical context shape interpretation.



7.4 Animation data dashboard

Through the dashboard built-in power BI, we can more intuitively see the specific situation of different genres of anime, including the number of anime average score studios and so on, through these indicators can analyze today's anime trends and make some predictions, using this dashboard can give anime production companies to bring detailed data so that they can carry out the subsequent production plan of anime.



8. Conclusion

In summary, our project has delved deep into anime data, revealing key trends and insights for industry stakeholders. By analyzing user behavior, ratings, and genre preferences, we've provided actionable intelligence to enhance content creation and marketing strategies. Leveraging Azure cloud services, we've streamlined data processing and visualization, ensuring efficiency throughout the workflow. Our findings underscore the importance of data-driven decision-making in optimizing the anime viewing experience and driving industry growth.

References

[1] Find architecture diagrams and technology descriptions for reference architectures, real-world examples of cloud architectures, and solution ideas for common workloads on Azure. **https://learn.microsoft.com/en-us/azure/architecture/browse/**

[2] Olympic Data Analytics | Azure End-To-End Data Engineering Project. **https://www.youtube.com/watch?v=laA9YNlg5hM&t=4515s**