NUS
National University
of Singapore

Analysis and Prediction of Video Games Sales across Regions

This project aims to predict video game sales across different regions and examine key factors that affect sales in each region. This was done by extracting common game names and selecting important features by implementing an ensemble of feature selections. XGBoost models were then fitted to predict both sales and log of sales in each region. One additional stacked model was then built by averaging the 2 model predictions. The result shows that in all four regions, the stacked models perform the best in terms of MSE and Adjusted R2.

A0219735X

1 Problem Introduction

The video game industry has experienced remarkable growth in recent years, establishing itself as a prominent form of entertainment and a multi-billion dollar global market. To optimize product targeting and revenue generation, it is crucial for game developers, publishers, and marketers to have a deep understanding of the local market dynamics that influence video game sales. This report will leverage advanced statistical methods and machine learning techniques to conduct a comprehensive analysis and develop predictive models for video game sales in different regions.

2 Dataset Description

The dataset for this project was retrieved from Kaggle. The data was scraped from VGChartz game database and Metacritic in Dec 2016. Metacritic was used to acquire game ratings, while VGChartz was used to obtain sales and other game data. This dataset has 16719 games and 16 variables: Name, Platform, Year_of_Release, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, Critic_Score, Critic_Count, User_Score, User_Count, Developer, and Rating. NA_Sales, EU_Sales, JP_Sales, and Other_Sales which indicate game sales in North America, European Union, Japan, and other regions, were used as response variables.

3 Data Cleaning and Exploratory Data Analysis

Since there are a lot of missing data (40 - 50%) in some of the variables, we only keep the rows with no null values, leaving 6825 rows. From Figure 1, we can see that most regions have 75% of sales value within a very small range. However, the range of sales across regions is not similar. NA and EU have a wider sales range, while Other and Japan have a narrower sales range. Hence, we suspect that model for NA and EU will perform worse than Japan and Other. Looking at Figure 4, 5, and 6, we also see that different regions have different popular genres, platforms, and publishers. Especially for Genre, we see that Japan has a different pattern

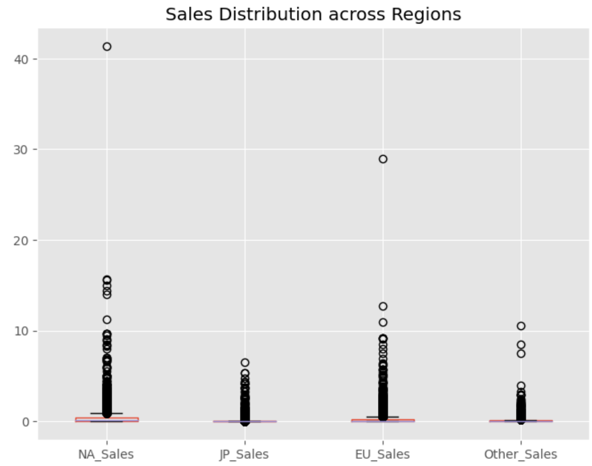

Figure 1: Sales Distribution Across Regions

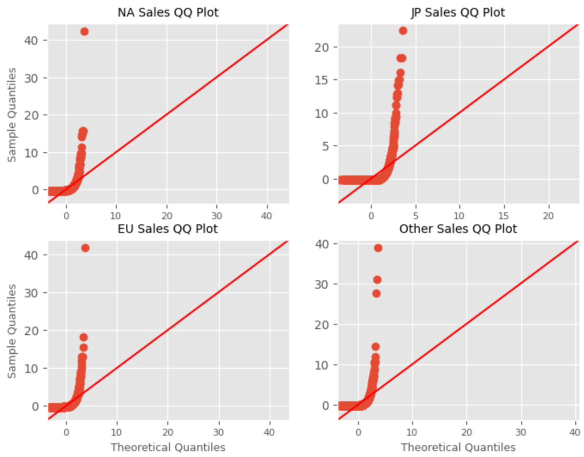

Figure 2: QQ Plots of Sales

where Role-Playing genre is the most popular which is not the case for the other 3 regions. Lastly, as can be seen from Figure 1 and 2, the distribution of sales is heavily right-skewed. However, as plotted in Figure 3, the distribution of log sales is ”closer” to normal. Hence, we would like to test whether models fitted on log sales perform better than on normal sales.

4 Feature Engineering and Extraction

Firstly, we generated the total year each game has been on the market by subtracting release year from 2017 since we suspect that total year has a positive impact on sales. 2017 is used since the data was scraped in late December 2016. Then, we split the data into 80% train and 20% test. After that, we extracted keywords from Name column of the train data by fitting CountVectorizer to it. We only took 1-gram till 4-gram that appear in at least 30 rows of the train dataset, which resulted in 165 features. These features are further reduced by removing 1-gram numbers, and n-grams that are a subset of other features (e.g. Harry $\subset$ Harry Potter). In the end, we have only 91 name keyword features. Subsequently, we one-hot encoded the categorical variables and scaled all predictors to range [0, 1]. Lastly, we split the data into 4 regions data (NA, JP, EU, Other), where for each region, we don’t use other region or global sales data as we won’t be able to look into future sales data when we are predicting it.

5 Feature Selection

To reduce dimension, we performed feature selection for each of the 4 regions dataset, consisting of Variance Threshold and an ensemble of 3 features selections to select 50 predictors (Forward Selection, Recursive Feature Elimination with Gradient Boosting Regressor, and Select K Best with Mutual Info Regression) similar to the feature selection process in C4 group project. After choosing variables selected by at least 2 out of 3 methods, 34 predictors were selected for NA region, 31 for JP, 32 for EU, and 29 for Other region. Each region has different selected publishers, developers, genres, and name keywords. All numeric variables Critic_Score, Critic_Count, User_Score, User_Count, and Total_Year are selected for all 4 regions except for User_Score which is not selected for JP region.

6 Models

For each region, we fitted 2 XGBoost models using the selected variables. The 2 models were fitted on the normal sales and log sales. For log sales models, actual sales value 0 in train data was converted to 0.01 (smallest sales value $>$ 0) first before taking the log. Aside from the 2 models, one additional stacked model is built by averaging the 2 model results (after converting log sales prediction back to normal sales). This stacked model was built due to the strengths and weaknesses of the 2 XGBoost models which we will discuss in Section 7 Evaluation. Moreover, the results from each of the 3 models were then passed to function $f(x) = max(0, x)$ to produce the final predictions as sales shouldn't be below 0. Lastly, note that all hyperparameters were tuned using GridSearchCV.

7 Evaluation

To fairly compare model performances, we converted the log sales prediction to normal sales before computing metrics. Also, due to percentage error calculation limitations, we converted the actual values 0 to 0.01 (smallest sales value $>$ 0) for MAPE computation.

Region	Model	MSE	MAPE	Adj R2
NA	Normal	0.364	229.30%	51.7%
	Log	0.404	111.61%	46.4%
	Stacked	0.364	168.22%	51.7%
JP	Normal	0.038	163.78%	25.6%
	Log	0.043	49.77%	17.0%
	Stacked	0.038	108.31%	25.8%
EU	Normal	0.187	323.37%	44.2%
	Log	0.192	121.48%	42.7%
	Stacked	0.172	219.33%	48.7%
Other	Normal	0.021	153.37%	47.12%
	Log	0.023	66.56%	43.16%
	Stacked	0.020	105.84%	49.98%

Table 1: Model Performances

From Table 2, we can see that for all regions in terms of MSE and Adj R2, Normal models are better than Log models. However, the MAPE for Log models is much lower than for Normal models. This implies the errors for small sales values in Normal models are much higher than in Log models and hence give higher penalties for MAPE. To combine the strength of these 2 models, we built stacked model that is just simply the mean of the 2 model predictions. Unsurprisingly, the MAPE for stacked models is better than Normal and worse than Log models. Surprisingly, the MSE and Adj R2 of the stacked models are very similar (for NA and JP) or even much better (for EU and Other) than the Normal models. Hence, we can use stacked models as our final models for all 4 regions.

8 Learnings

To learn predictors' effect on sales, we generated the SHAP values for Normal and Log models for all regions. Then, we examined predictors' effects using both models' SHAP values. For all regions, User Count, User Score, Critic Count, Critic Score, and Total Year have positive effects on sales which are expected. Additionally, some notable and interesting predictors' effects are listed in Table 2.

Region	Positive	Negative
NA	Platform Wii Developer Nintendo Publisher Electronic Arts Genres: Sports, Fighting Rating E Games: Mario, Call of Duty	Platform PC
JP	Platforms: DS, PS3 Developer Nintendo Publishers: Nintendo, Namco Bandai Games Genre Role-Playing Games: Mario, Pro Evaluation Soccer	Platforms: PC, WiiU
EU	Platforms: Wii, PS2 Developer Nintendo Publisher Nintendo Rating E Games: Lego, FIFA Soccer, Call of Duty	Platform PC Genre Role Playing
Other	Platforms: PS2, PS3, Wii Developer Nintendo Rating E Games: Lego, FIFA Soccer, Call of Duty	Platforms: PC, WiiU, GC Genre Role-Playing

Table 2: Positive and Negative Features across Regions

9 Recommendations & Conclusions

From Table 2, some recommendations for game developers and publishers are:

1. If you wanna develop video games, develop games suitable for every age (age rating E) and choose other platforms than PC, e.g. Wii, PS2, or PS3

2. Games published or developed by Nintendo are popular across regions. If you are an indie/solo developer, you may want to publish your games on Nintendo e-Shop which can be done by following this instructions.

3. Games such as Mario (Bros, Kart, etc), Lego, FIFA Soccer, and Call of Duty are popular in most regions. Hence, releasing the sequel/variations of them might be beneficial.

4. If your target market is JP, you can develop Role-Playing games, but not for EU and Other region. For NA, you can develop Sports/Fighting games

Hopefully, with this project, game developers and publishers are able to gain deeper insights into the market and make data-driven decisions to guide business decisions.

10 Appendix

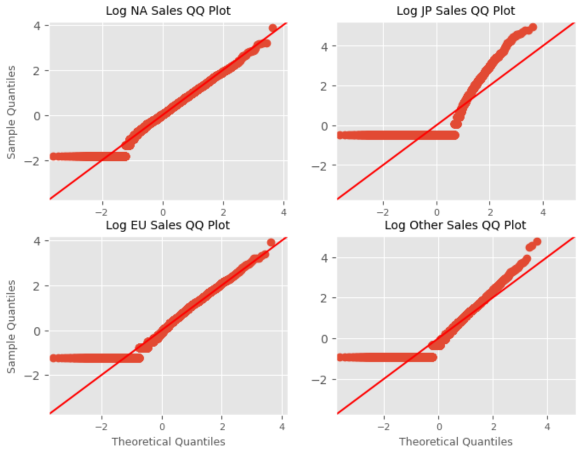

Figure 3: QQ Plots of Log Sales

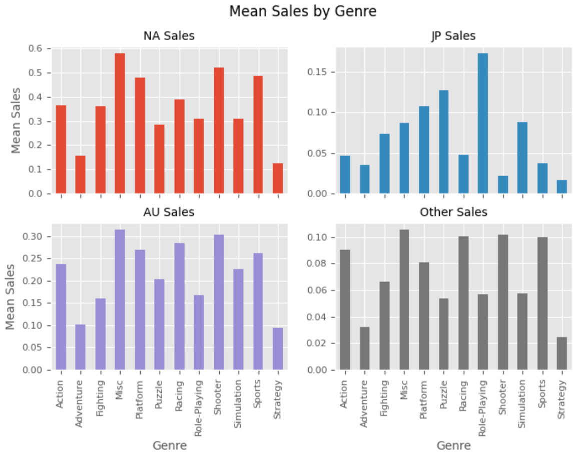

Figure 4: Mean Sales by Genre

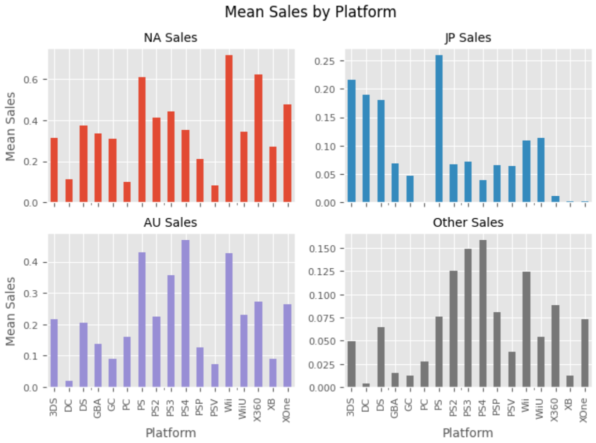

Figure 5: Mean Sales by Platform

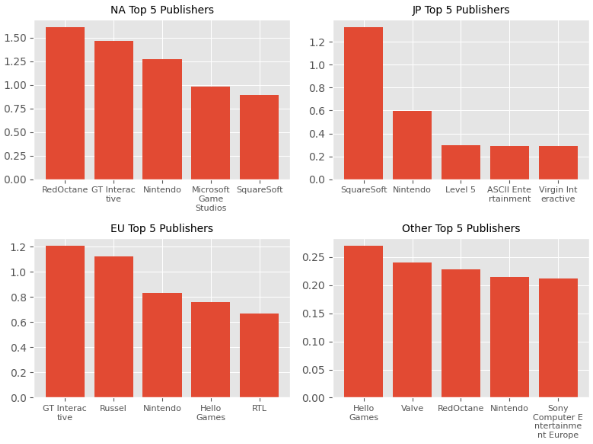

Figure 6: Top 5 Publishers