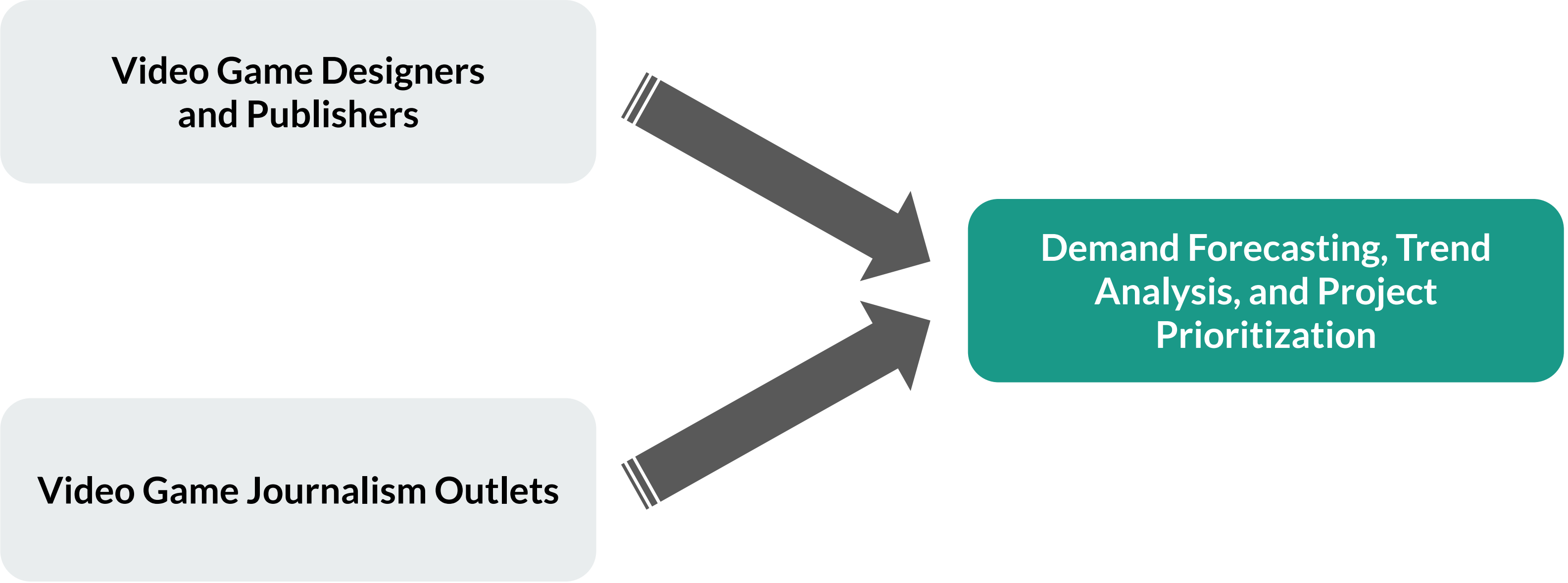
Video Game Sales Data Analysis

Adam Young

QUESTION:

Can we predict the global sales of a video game using only it's basic characteristics?

Who Cares?



Video Game Designers and Publishers

Video Game Journalism Outlets

Demand Forecasting, Trend Analysis, and Project Prioritization

Data Description

Observations

16,598 Individual Video Games

Source

Web scraped from vgchartz.com

Descriptive Variables

Release Year

Publisher

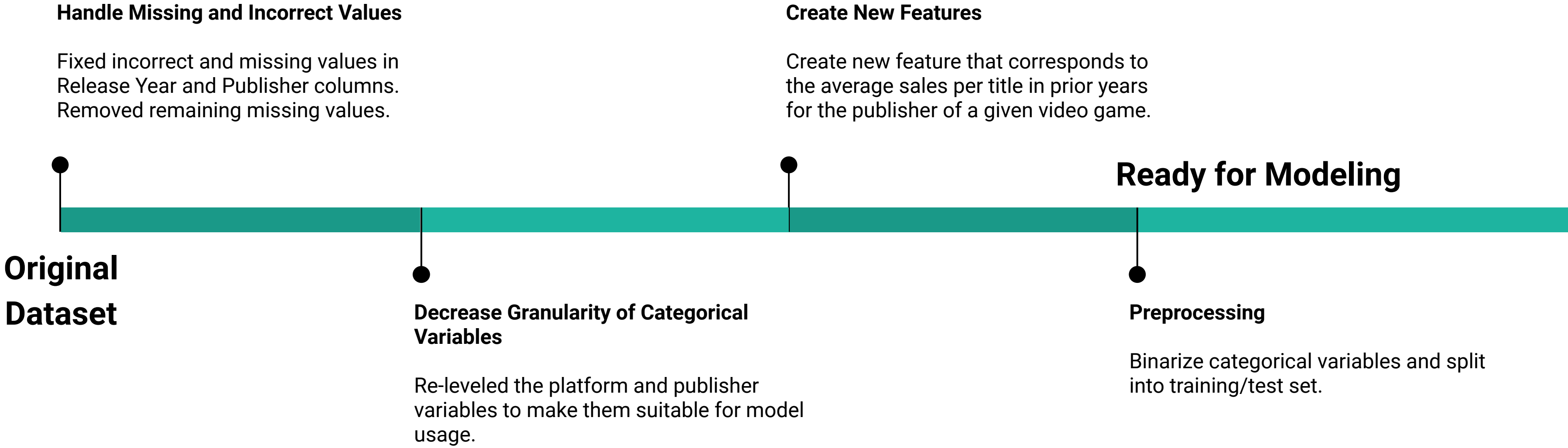
Platform

Genre

Target Variable

Global Sales

Data Cleaning, Feature Engineering, Preprocessing



Original Dataset

Handle Missing and Incorrect Values

Fixed incorrect and missing values in Release Year and Publisher columns. Removed remaining missing values.

Decrease Granularity of Categorical Variables

Re-leveled the platform and publisher variables to make them suitable for model usage.

Create New Features

Create new feature that corresponds to the average sales per title in prior years for the publisher of a given video game.

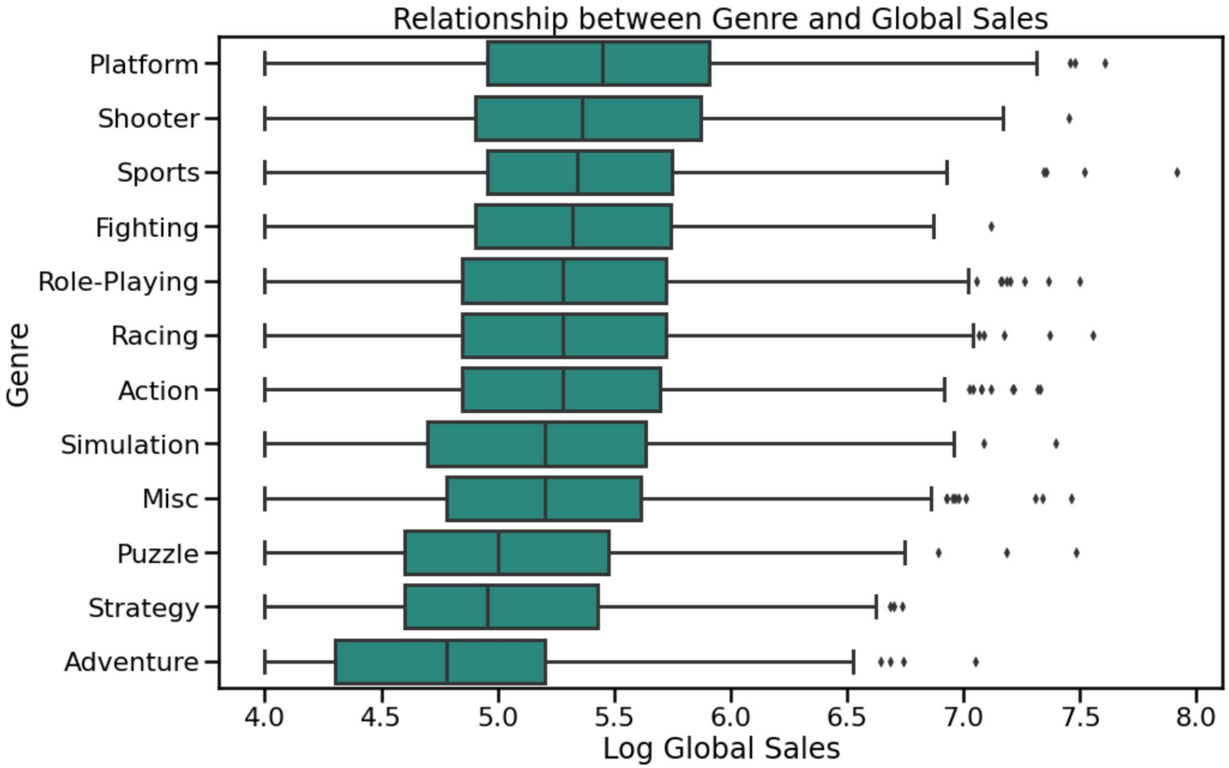
Preprocessing

Binarize categorical variables and split into training/test set.

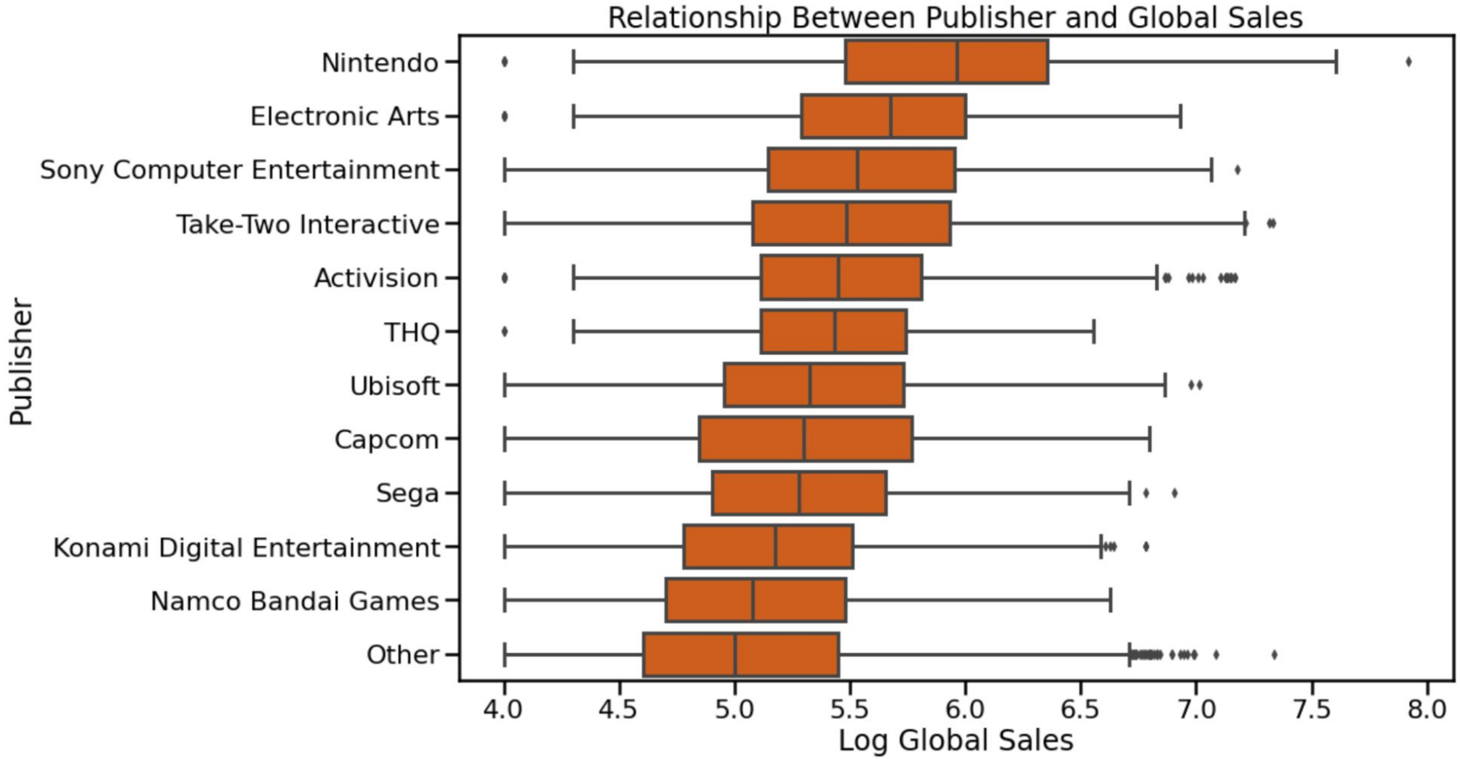
Ready for Modeling

Exploratory Data Analysis

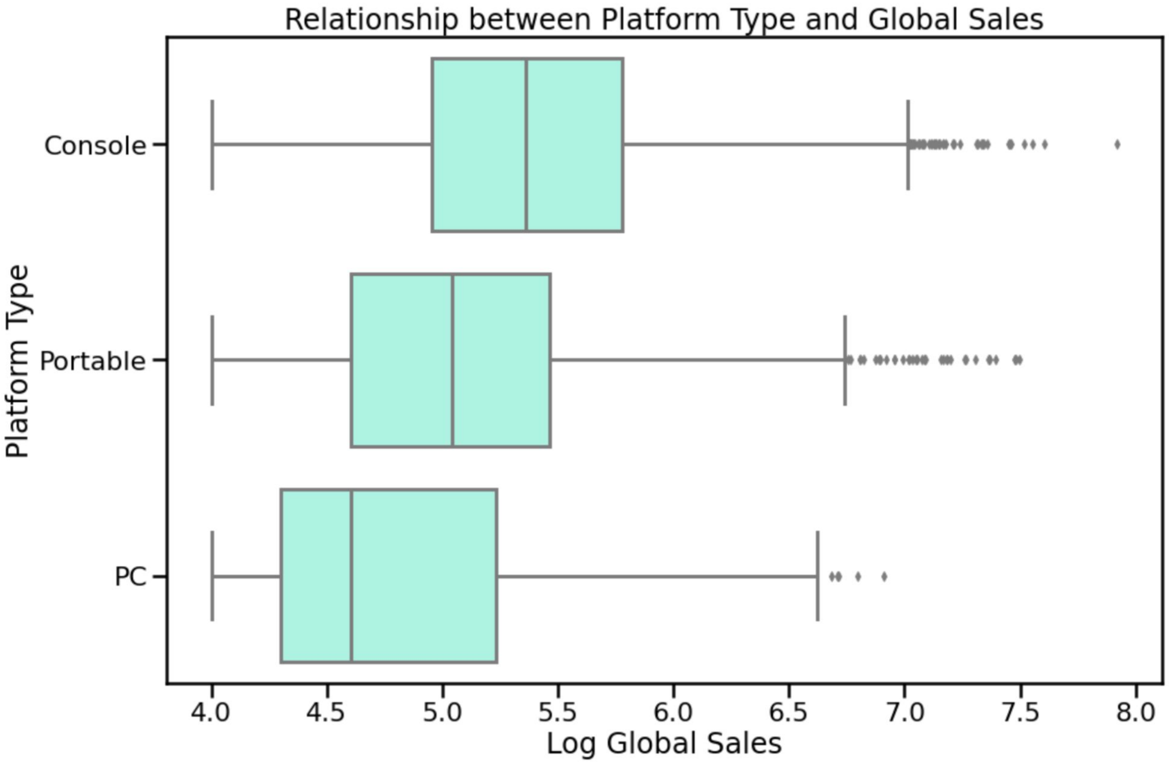
Global Sales by Genre



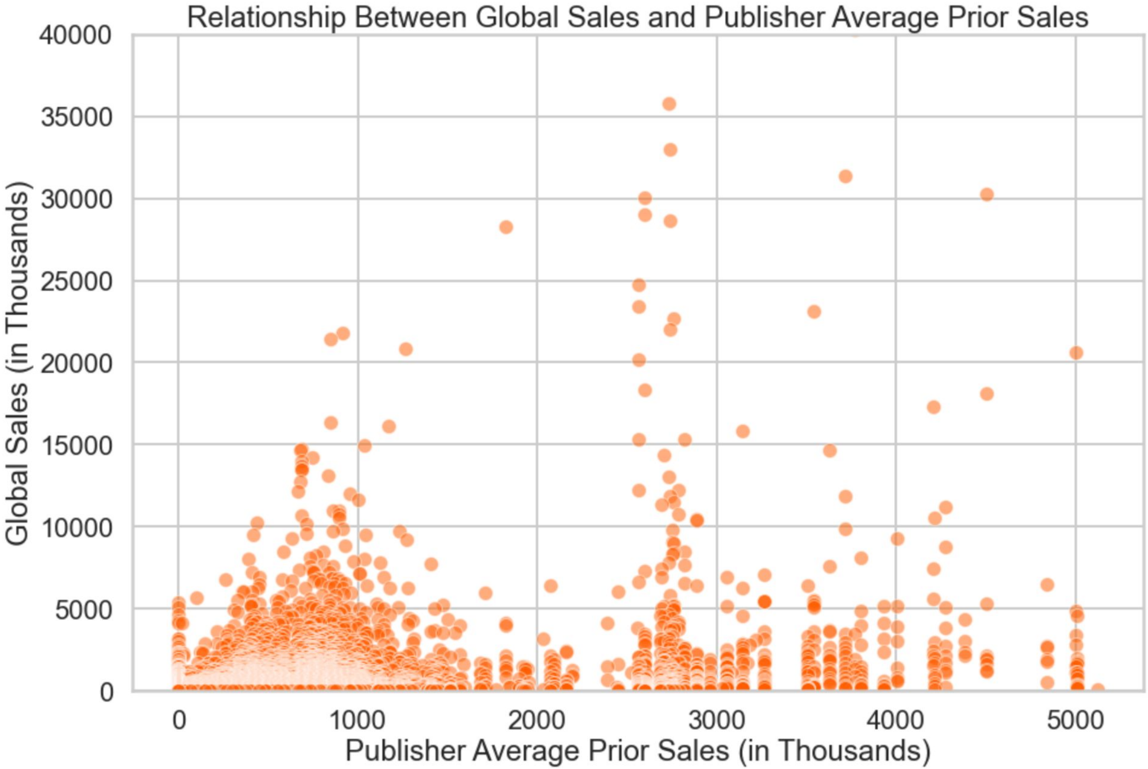
Global Sales by Publisher



Global Sales by Platform Type



Global Sales and Publisher Average Prior Sales



Pre-Modeling Concerns

01

Small number of available predictor variables

- 4 variables in starting dataset
- Small number of possible relationships with Global Sales

02

Only one numerical feature

- One numerical feature, created during the feature engineering stage
- Difficult to predict a continuous target with only a few categorical features

03

High variance within categories with respect to Global Sales

- Evident in the wide whiskers on the previous boxplots
- Makes the categorical variables even less useful for predicting a continuous target

NEW QUESTION:

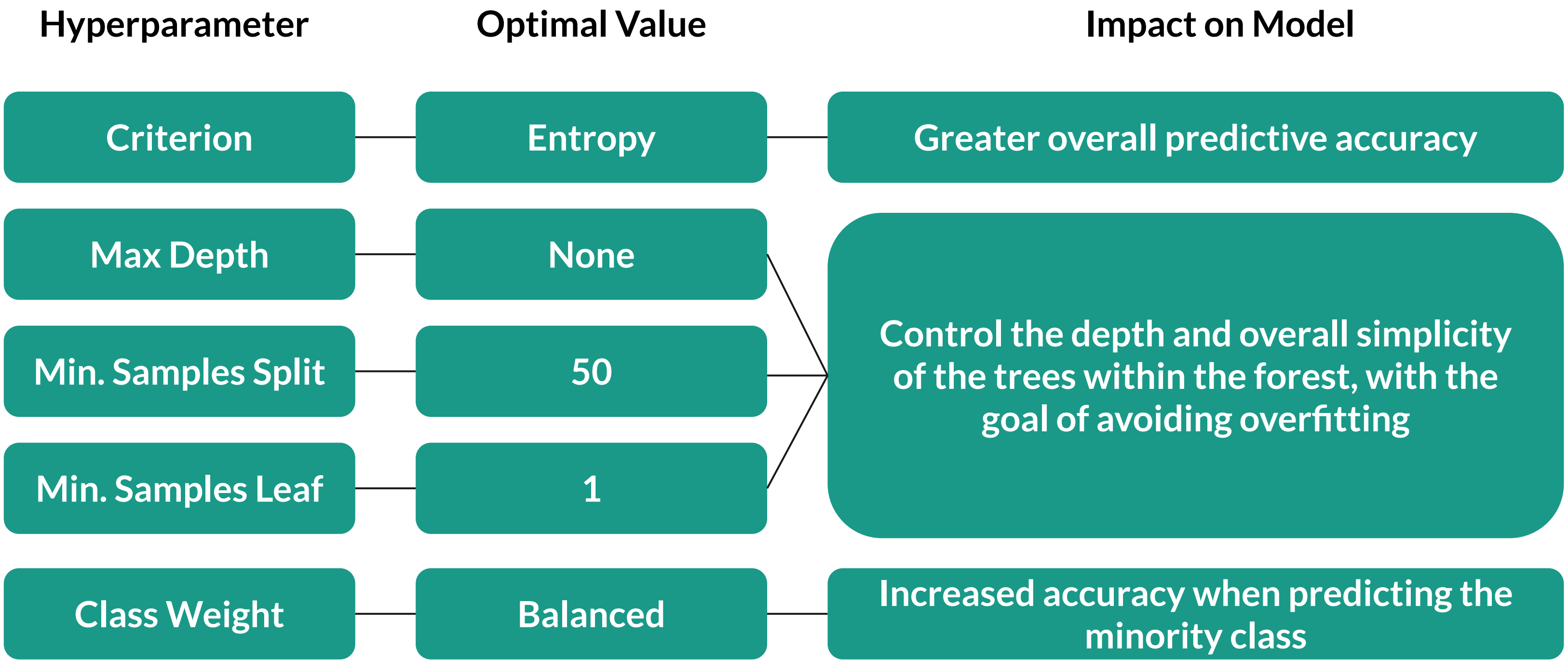
Can we predict whether a video game will sell >500,000 copies globally using only it's basic characteristics?

Modeling and Results

Baseline Modeling

	Training Error	Test Error	Potential for Improvement
Linear Regression	🔴	🔴	🔴
Random Forest Regression	🟠	🔴	🟠
Logistic Regression	🔴	🔴	🔴
Random Forest Classifier	🟠	🟠	🟢

Random Forest Classifier Hyperparameter Tuning



Hyperparameter	Optimal Value	Impact on Model
Criterion	Entropy	Greater overall predictive accuracy
Max Depth	None	Control the depth and overall simplicity of the trees within the forest, with the goal of avoiding overfitting
Min. Samples Split	50	
Min. Samples Leaf	1	
Class Weight	Balanced	Increased accuracy when predicting the minority class

Final Model Metrics

Metric	Performance on Training Set	Performance on Test Set	Difference
Accuracy	0.772	0.739	-0.033
Precision	0.522	0.470	-0.052
Recall	0.795	0.721	-0.074
F1-Score	0.630	0.569	-0.061

Key Takeaways

A publisher's past success is a good indicator of future success.

Smaller publishers sell less than larger publishers.

Console > Portable > PC

'Adventure', 'Puzzle' and 'Strategy' are the lowest selling genres by a significant margin.

This model should not be used as the sole predictor of a game's success, though it may be helpful when combined with other more subjective methods.

Further Research

Are there any other quantifiable metrics that could help improve the predictive accuracy of this model? (ESRB Rating, Is the game part of a series?, etc.)

Explore creating text-based features (Does the title contain ‘Mario’, ‘Call of Duty’, etc.)

Experiment with different levels of categorical granularity. We used low categorical granularity partially due to limited processing power.