Video Game Sales Data Analysis and Classification

Problem Statement

The client (IGN) has tasked us with identifying whether an accurate sales prediction can be made using only basic characteristics about a video game. This model will be used to help IGN identify which articles should be prioritized and placed on the front page of their website, with the hopes of increasing clicks and ad revenue by 1-5% in the next fiscal year.

Data Cleaning and Feature Engineering

Since a dataset was not provided by the client, we have web-scraped a comprehensive list of video games from vgchartz.com. This list contains 16,000+ video games released between 1985 and 2017. The following fields were included for each video game in the dataset:

- Title
- Platform
- Release Year
- Genre
- Publisher
- North America Sales
- Europe Sales
- Japan Sales
- Other Sales
- Global Sales

Adjustments were made to correct missing and incorrect values in the Release Year and Publisher columns. Once the data was cleaned, the following additional features were generated:

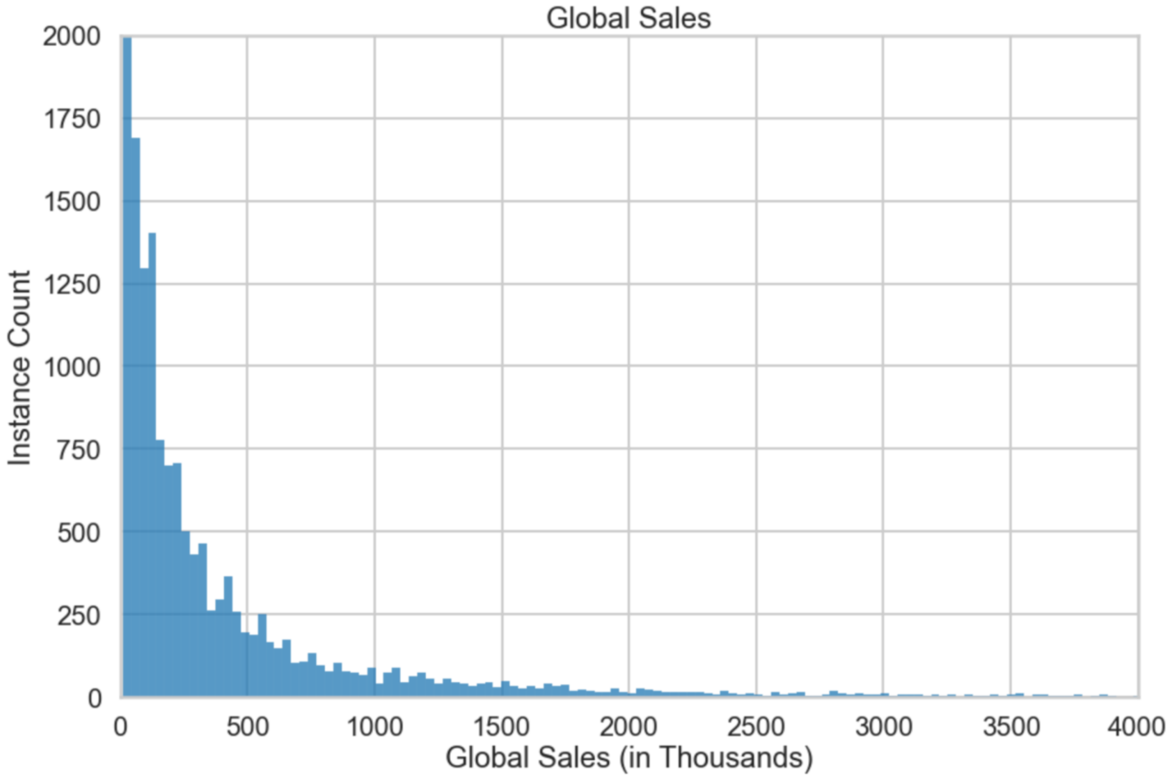
Feature Name	Feature Description
Platform Category	Relevel the initial Platform variable (which had 31 unique categories) into 5 distinct categories (PlayStation, Xbox, PC, Nintendo, Other).
Platform Type	Relevel the initial Platform variable (which had 31 unique categories) into 3 distinct categories (Console, Mobile, PC).
Publisher Title Count	Count of all titles released by a given publisher.
Publisher 2	Relevel the Publisher variable (which had 700+ unique categories) into 12 categories, including the 11 largest publishers along with an ‘Other’ category.

Publisher Average Prior Sales	Average sales per title released by the observation publisher prior to the observation year.

By reducing the granularity in some of the categorical variables, and creating two new numerical variables, we may be able to increase the predictive accuracy of our models.

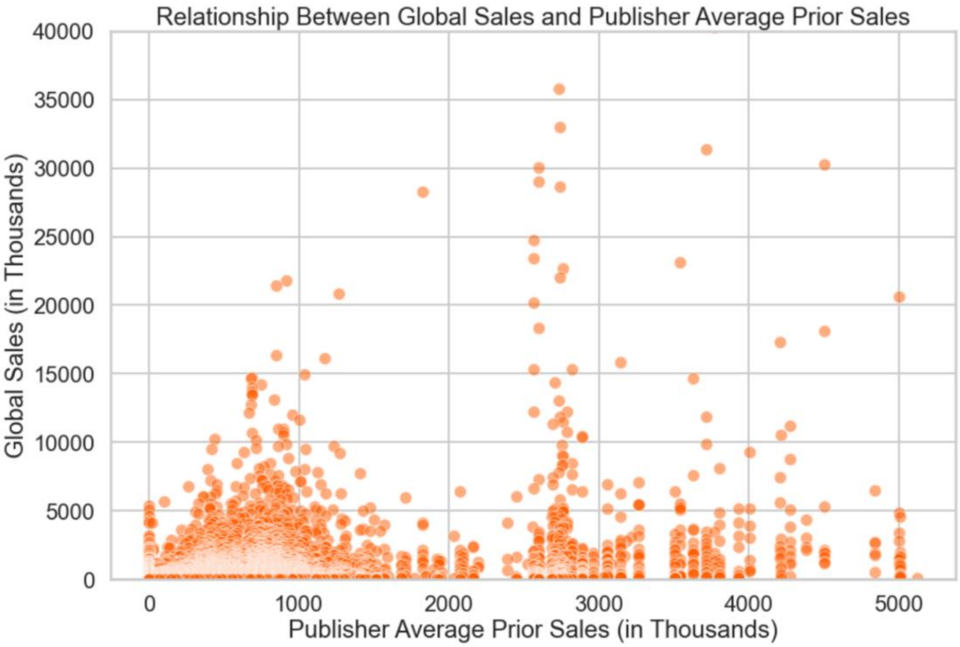
Exploratory Data Analysis

The clear choice for the target variable in this dataset is Global Sales since it is the sum of sales in all regions and signifies the overall demand for a video game. It has a highly right-skewed distribution, with most video games selling less than 500,000 copies. When training regression models we will log transform the Global Sales variable to shift it closer to a normal distribution.

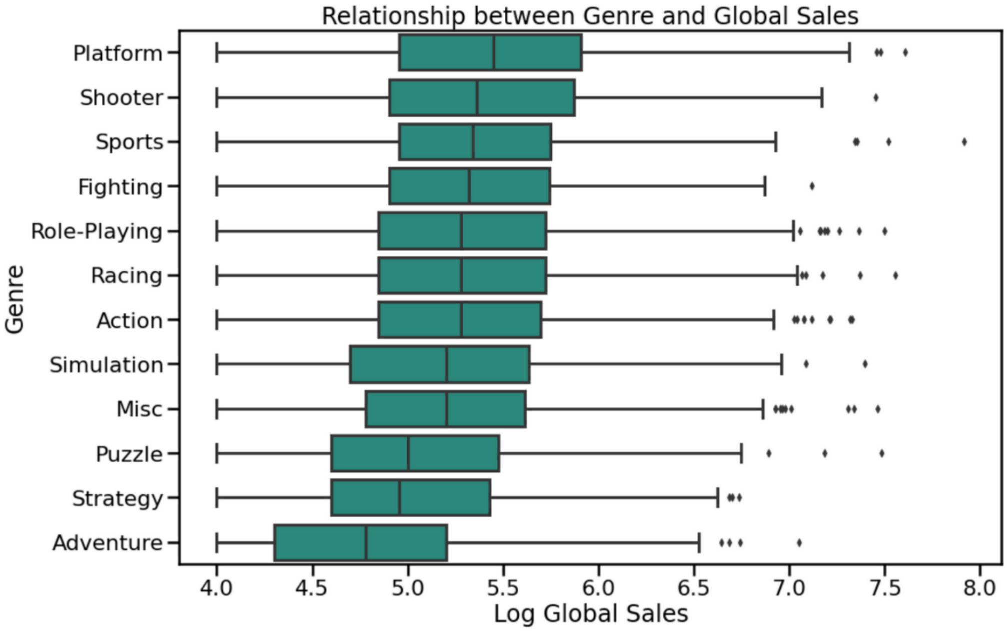


While exploring the relationships between Global Sales and the explanatory variables, there were four main relationships identified.

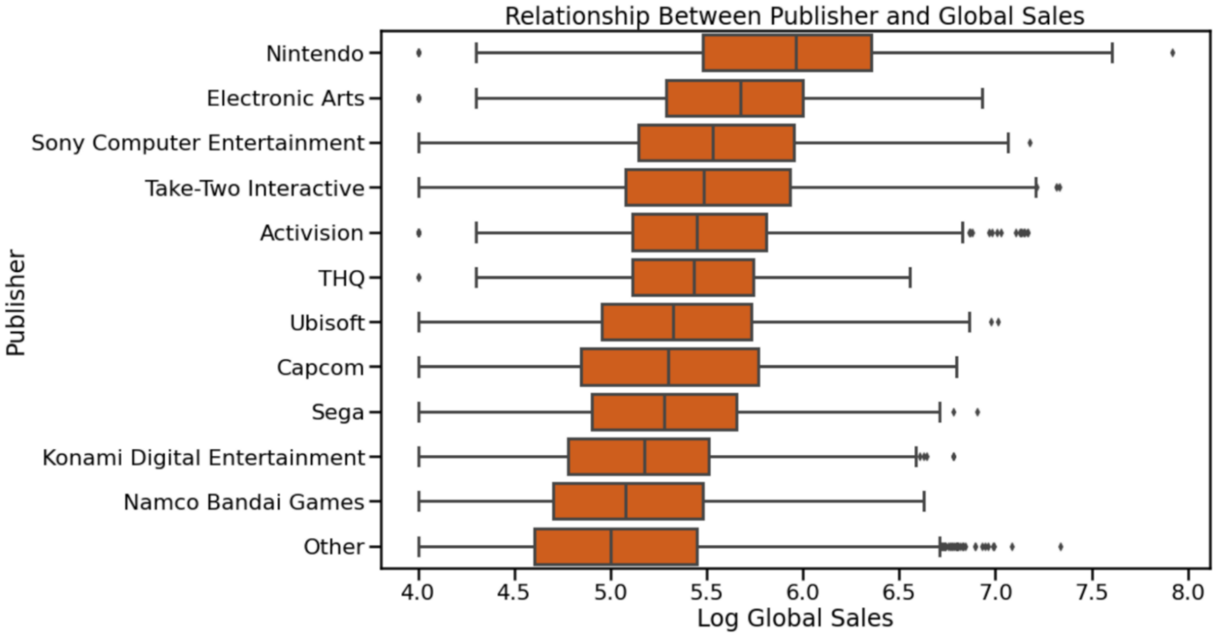
1. There is a slight positive correlation (~0.4) between the Global Sales for a given observation and its publisher's average sales per video game in prior years.



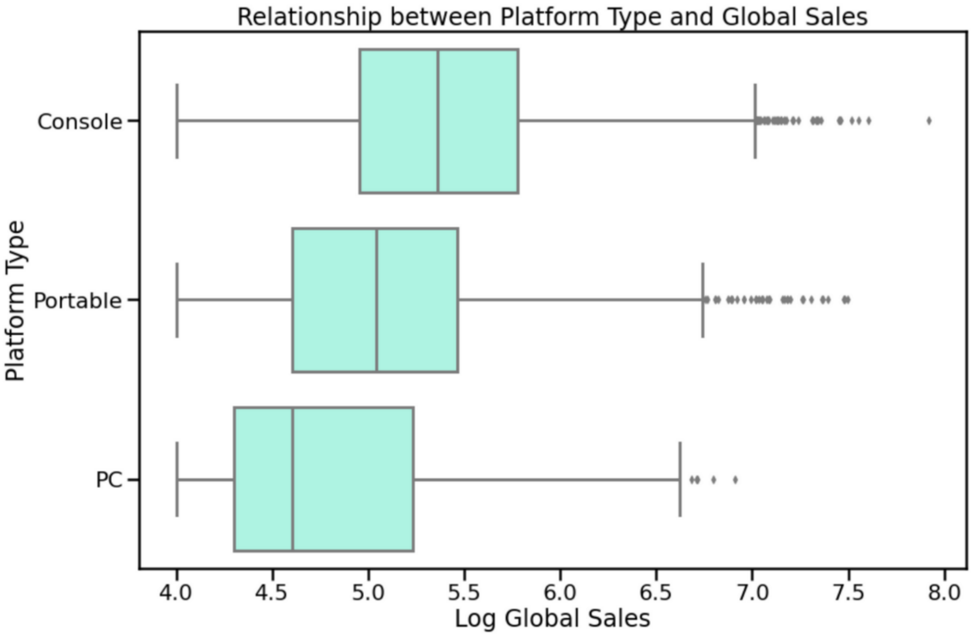
2. Some genres appear to have significantly lower median Global Sales values than others.



3. The largest publishers have significantly higher median Global Sales values than the smallest publishers.



4. Console games have the highest median Global Sales values, followed by portable games and then PC games.



While there are clear relationships between some of the explanatory variables and the target variable, this data set is unlikely to enable us to make accurate predictions of a continuous target. The first reason for this is that the dataset only includes one true numerical feature, which is the Publisher Average Prior Sales feature that we created. The other variables are either categorical or can be treated similarly to a categorical variable since they do not have a clear linear relationship (Release Year). The second reason is that there is a large degree of variance within each of the categories of the categorical variables, which can be seen in the wide whiskers on each of the boxplots above. The result of this high variance is that while we can identify general relationships between the categories, we can not translate that into accurate predictions of a continuous Global Sales variable. To solve this problem, we may have to transition to a classification approach and try to predict whether a video game will have Global Sales above or below a certain threshhold.

Baseline Modeling

To assess which type of model would best suit this data, I ran three baseline models and evaluated their performance before moving on to hyperparameter tuning.

Model	Performance
Linear Regression	R^2 = 0.295, MAE = 0.982
Random Forest Regressor	R^2 = 0.631, MAE = 0.666
Random Forest Classifier ($1M Threshhold)	Accuracy = 0.91, Precision = 0.83, Recall = 0.36

The linear regression model performed extremely poorly, as expected. The random forest regressor was a significant improvement on the linear regression, though still only had an R^2 value of 0.631. Both regressors suffered from overfitting and had greatly reduced performance when evaluated on the test set. The random forest classifier performed decently well, with 91% accuracy, however it did not predict the minority class well. The minority class is the most important class since it represents the video games with over 1M global sales.

The random forest classifier will be pursued further since it has the highest potential for increased performance through hyper parameter tuning and class balancing.

Model Tuning and Final Model Metrics

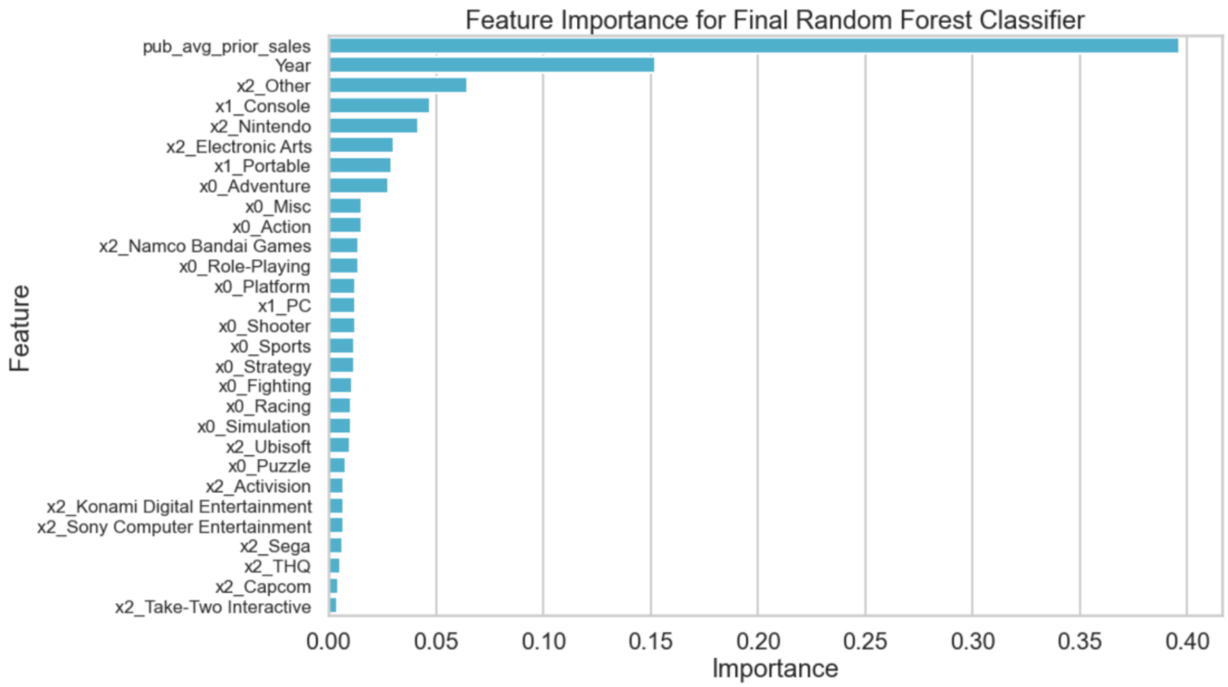
There are two methods that I decided to use to tackle the problem of poor prediction performance on the minority class. First, I changed the target variable to be an indicator of whether the video game sold over 500,000 copies instead of 1,000,000. This provided more support for the minority class, while not damaging the usefulness of the model by too much. Secondly, I performed a cross-validated grid-search to optimize the hyperparameters of the random forest classifier. One such hyperparameter was the Class Weights parameter, which allows us to put higher weights on the minority class so that the model places more emphasis on correctly predicting that class. The final random forest classifier hyperparameters can be found in the table below:

Hyperparameter	Final Value
Bootstrap	True
Ccp_alpha	0.0
Class_weight	Balanced
Criterion	Entropy
Max_depth	None
Max_features	Auto
Max_leaf_nodes	None
Max_samples	None
Min_impurity_decrease	0.0
Min_impurity_split	None
Min_samples_leaf	1
Min_samples_split	50
Min_weight_fraction_leaf	0.0
N_estimators	100
N_jobs	-1
Oob_score	False
Random_state	123
Verbose	0
Warm_start	False

The performance metrics for the final random forest classifier can be seen below for both the training set and the test set. While there was still a drop-off in performance between the training and test set, the magnitude of the drop-off is significantly less than both the regression models and the non-tuned classifier model. Most importantly, through hyperparameter tuning and changing the classification threshhold, we were able to improve the recall from 0.36 to 0.72.

Metric	Training Set	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

The feature importance chart for the final random forest classifier can be seen below. The most important feature by far was the average prior sales for each publisher. Other useful features were Year, Console Type, and Publisher (specifically the Other publisher category, and Nintendo).



Recommendations and Further Research

While the hyperparameter tuning significantly improved the performance of our classifier, its low precision (0.470) and recall (0.721) mean that there are many errors, including both false positives and false negatives. We would not recommend using this model as the sole predictor of video game success, but some of its insights may be helpful when combined with other non-quantifiable knowledge that the model doesn't have access to, such as customer hype levels, early-access reviews, and gameplay previews. The primary insights that are worth pulling from this model are as follows:

- Publishers with higher average sales in prior years are more likely to have a highly selling new release than publishers with lower average prior sales.
- Smaller publishers are less likely to have a highly selling new release than larger publishers.
- Console titles are likely to have the highest sales, followed by mobile titles and then PC titles.
- 'Adventure' is the lowest selling genre by a significant margin.

Further Research Recommendations:

- Are there other quantifiable metrics available for use? (ESRB Rating, Is the video game part of a series, etc.)
- Explore creating text-based features (i.e., does the title contain 'Mario' or 'Call of Duty', etc.).

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