Nate Lewis

STAT 486

17 April 2024

Video Game Sales Prediction: Report

Introduction

Video games are a multi-hundred-billion dollar a year industry. They are also a hobby of mine. With this project I hope to see if it is possible to predict how much video games sell using machine learning tools. In the process of doing this I was particularly intrigued to see which of my features would have the biggest influence on predictions. Is quality king or does the resources at your disposable determine your success?

EDA

I used 10 features when making my predictions for North American sales numbers. The 10 features were release year, critic score, number of critics, user score, number of users, platform, sales in Japan, genre, publisher, and rating. Below are the summary statistics for my data set. The numerical features are on the left and the categorical features are on the right.

	Year_of_Release	NA_Sales	JP_Sales	Other_Sales
count	16719.000000	16719.000000	16719.000000	16719.000000
mean	2006.495604	0.263330	0.077602	0.047332
std	5.831862	0.813514	0.308818	0.186710
min	1980.000000	0.000000	0.000000	0.000000
25%	2003.000000	0.000000	0.000000	0.000000
50%	2007.000000	0.080000	0.000000	0.010000
75%	2010.000000	0.240000	0.040000	0.030000
max	2020.000000	41.360000	10.220000	10.570000

	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Count
count	16719.000000	16719.000000	16719.000000	16719.00000	16719.000000
mean	0.533543	70.010886	23.609068	7.32978	86.752856
std	1.547935	9.776336	13.509406	1.02773	384.376835
min	0.010000	13.000000	3.000000	0.00000	4.000000
25%	0.060000	71.000000	21.000000	7.50000	24.000000
50%	0.170000	71.000000	21.000000	7.50000	24.000000
75%	0.470000	71.000000	21.000000	7.50000	24.000000
max	82.530000	98.000000	113.000000	9.70000	10665.000000

	Platform	Genre	Publisher	Developer	Rating
count	16719	16719	16719	10096	16719
unique	31	12	581	1696	8
top	PS2	Action	Electronic Arts	Ubisoft	E
freq	2161	3372	1410	204	10760

A few take aways from these statistics include the increase in the number of games produced every year, the fact that North America has greater sales than other regions, and that users and critics tend to give games similar scores, with critics being slightly more critical.

Below are a few of the plots I created during the EDA Process:

On the top left is the distribution of video game sales in North America. It is extremely right skewed, so I decided to apply a log transformation in order to mitigate some of the effects of this skewedness. The remaining three plots show the relationships between the transformed log of North American sales with genre, release year, and critics scores. These plots along with other similar ones I created show that while there is certainly a relationship to be seen between sales and the features I'm using, none of them come close to being truly predictive on their own.

Methods

Feature engineering was pretty minimal for this project. As previously mentioned, I did transform North American sales using a log function due to the skewedness of the data. I also did

dimension reduction using both PCA and TruncatedSVD to reduce the number of components down to two, to experiment with the results that came from that.

I experimented with the following five models to make predictions on my data:

1. Linear Regression: Linear Regression is a straightforward statistical method that models the relationship between a dependent variable and one or more independent variables using a linear equation. There were no hyperparameters to explore with this model. Using this model I achieved an MSE: of 0.05, the highest of all the models I tried.
2. Decision Trees: Decision Trees are a type of model that uses a branching method to illustrate several possible outcomes of a decision, which can be visualized like a tree. For this model I explored different values for max_depth, min_samples_split, and max_features. My best model of this type had an MSE of 0.04.
3. Random Forests: Random Forests build multiple decision trees and merge them together to get a better prediction. With this model I tested different values of n_estimators, max_features, and max_depth. My best Random Forests had an MSE of 0.03.
4. Gradient Boosting: Gradient Boosting is an advanced ensemble technique that builds trees one at a time, where each new tree helps to correct errors made by previously trained trees. For hyperparameter tuning I tried different values of n_estimators, learning_rate, and max_depth. My best model of this type had an MSE of 0.03.
5. Neural Network: This model consists of layers of “neurons”, designed to approximate any function representing data relationships. I tested different values for the hyperparameters batch_size, learning_rate, and number of layers and neurons per layer. My best MSE with this model was 0.041.

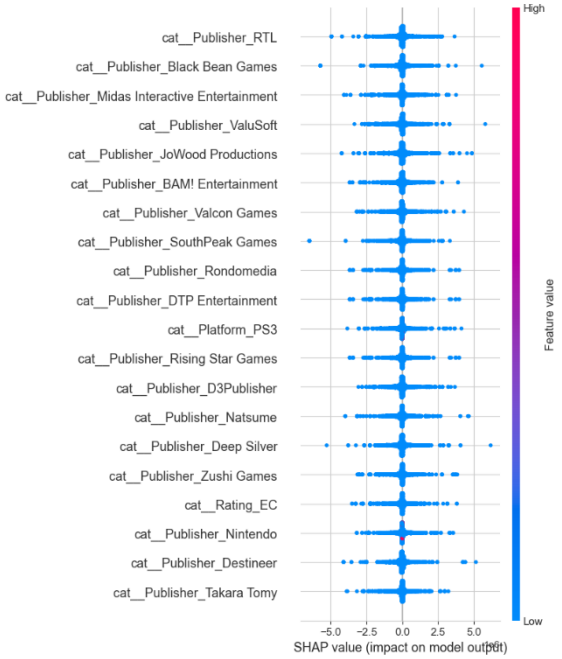
Discussion of Model Selection

I decided on which models to attempt to use based on my knowledge of which models perform well with regression tasks and with the goal of using models of diverse types even if I didn't think some of them would work as well as other options. Overall, the ensemble models performed better than single models. I think decision tree models and linear regression are just a little too simple to work as well as the ensemble models do. On the other hand, I think the neural network didn't work as well as due to issues with overfitting. I don't think this dataset required the kind of computational complexity that neural networks bring.

Detailed Discussion on Best Model

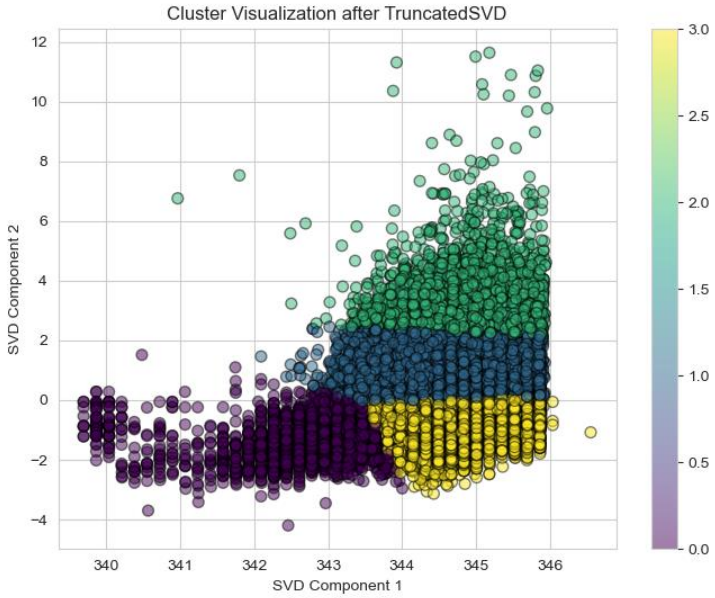
The model that worked best for me was the random forests model. I think this is due to this model's protection against overfitting and how well suited it is for regression problems with features of different types. I used RandomizedSearchCV for the hyperparameter tuning, though in the end it was the original model with the default values that happened to perform the best. This model had an MSE of 0.03 and an R^2 of 0.61.

I used SHAP to determine which features were most important for these predictions resulting in the following plot:



This shows that generally the publisher of the game is the biggest predictor. This makes sense as the publisher would be heavily correlated with the marketing budget and public awareness of a game. While there are no false positives or the like for this case, I did compute residuals to determine what the biggest outliers were. The titles with the biggest positive residuals included *Wii Sports*, *Just Dance 3*, and *Call of Duty: Black Ops 3*. The titles with the most negative residuals included *PaRappa the Rapper*, *Bayonetta*, and *Tomodachi Life*.

I also did cluster analysis. Using the elbow method I determined to make 4 clusters and then using TruncatedSVD dimension reduction I made the following plot:



The data fit very well into these clusters. After some analysis I determined that the purple cluster was connected to retro games, the yellow cluster was connected to commercial and critical failures, the blue cluster represented games that were either hit indie games or mild successes from big publishers and the green cluster was the big hit games.

Additional Method

I also used the apriori algorithm from the mlxtend library to do association rule mining. The basic goal of this method is to identify frequent or significant patterns between categorical variables. This shows you which combinations of features give unique results. To do this I also transformed user score and critic score into categorical variables by dividing scores into low, mid, high, and very high groups. The results showed that critically acclaimed shooters with M ratings, sports games that are Rated E with low user scores, and E rated games published by Sony were among the combinations of features that performed the best in sales.

Conclusion and Next Steps

In conclusion, the utilization of Random Forests proved to be the most effective model for predicting video game sales. This highlights how ensemble techniques like this one are often good at modeling the complexity of regression tasks with diverse feature sets. Moreover, the analysis highlighted the large role of publishers in determining sales figures, confirming the importance of marketing efforts and public awareness over all else when it comes to financial success. Moving forward, potential improvements could involve exploring hybrid models that combine the strengths of different algorithms or doing more feature engineering to uncover additional insights.