Video Game Sales

Global sales prediction

Dan Heikenberg

EC Utbildning

Examination- Project I Data Science

202410

Abstract

This report is in two parts. In the first part we try to explore and clean data related to video game sales, in Jupyter Notebook with any necessary plugins, and find a machine learning model suitable for training with this data, to make a model that can be used to predict global video game sales. The data, found at Kaggle.com - a machine learning and data science community, includes columns with data such as video game names, publishers, user scores, and global sales. Here we also try to explain the output (global sales) of the chosen machine learning model. After finding great challenges during the exploratory data analysis, and trying different models, we were able to work our way to a better model. The model is not good enough, but there´s room for improvements.

The second part is an attempt to highlight some of the more interesting parts of the data, including the feature(s) that explains the output of the chosen model. This is done with both SHAP and power bi.

Contents

1 Introduction

According to Statista (Statista, n.d), the global video game industry is and has been a billion-dollar business for many years. The revenue of the worldwide gaming market in 2022, was estimated at almost 347 billion U.S. dollars, and the mobile gaming market had an estimation of 248 billion U.S. dollars of the total. Therefore, it might be interesting for companies, publishers, game developers to have more insight in what games create the most revenue. It could be genre trends that change from year to year, or certain publishers that somehow (maybe marketing strategies) makes a game more successful.

With data over video game sales, it would therefore be interesting to see if one could train a model that could help predict global sales, and what features, if any, explains the output. That is the purpose of this rapport, and to create a starting point for further research, data gathering and model training.

1. Is it possible to use this dataset to create a model predicting global sales?
2. Can we find the feature(s) explaining the global sales (the output of the model)?

2 Theory

2.1 Single and multiple Linear Regression Modelling

Linear regression is a model that estimates a linear relationship between a response and a single, or multiple explanatory variables/predictors. The price of a product can be the response variable, and size, material and so on, can be the predictor variables.

2.2 Decision Tree and CatBoostRegressor Modelling

Instead of each input continuously resulting in a different output in linear regression, decision tree regression takes into account that there might be only one output for a set of inputs. The prediction of linear regression model is continuous in nature (Vitaflux, n.d). Linear regression models have a parametric mathematical equation consisting of one or more parameters. Decision tree regression models are non-parametric, and only need decision about which column or feature to choose for each node of the tree and what value to split the data on that node (Vitaflux, n.d)..

CatBoost is a type of decision tree model. It has an ability to integrate a variety of different data types (Medium, n.d).. It also offers a way to handle categorical data, requiring a minimum of categorical feature transformation. Many other models can´t handle non-numeric data from the go.

2.3 MSE, MAE, R-squared and RMSE

The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set.

The R-squared value is the proportion of the variance in the response variable that can be explained by the predictor variables in the model (Statology, n.d.).

The root mean square error tells us the average distance between the predicted values from the model and the actual values in the dataset. This shows how much error on average the predictions have.

The lower value of MAE, MSE, and RMSE implies higher accuracy of a regression model. However, a higher value of R square is considered desirable (Medium, n.d).

2.4 SHAP

SHAP is a tool that helps to interpret machine learning models, how trained models with their features impact the predictions (Datacamp, n.d). It works with any machine learning model, and may be used to show, for example, the importance of features in a model, and if more/less counts of a feature impacts the model negatively/positively.

3 Method

Video game data was fetched from Kaggle. This data was imported and explored in Jupyter Notebook. It was then trained with a linear regression models, both with a transformed data set, and a data set where missing rows had been removed. This was made as a decision after the exploratory data analysis, where we discovered problems in the data in the form of large quantities of missing values.

After comparing the results of the models, we tried a decision tree model to see if we could get better results. With this model, we used SHAP to explain the feature importance of this trained model.

Lastly, a power bi report was made, which amongst other things further highlights the discovered feature importance of the trained decision tree model.

3.1 Data collecting

Data concerning video game sales was fetched from Kaggle.com - a machine learning and data science community (Kaggle, n.d). Here data can be downloaded to be used for data modelling. The data was chosen by searching their website for “video games” and choosing a data set without looking any further at it.

3.2 Agile methodology

We have worked with agile methodology. Instead of doing all the work independently, we have had continuous contact, discussions, ideas, planning, and we have helped and supported each other on every step before continuing to the next step together, while leaving the door open for changes in the strategy while going forward. This has kept us motivated, focused, reflecting and aware of each other’s progress which gives a better whole perspective and lowers stress. It has also helped checking and deliver working and tested code during the journey, and made us a flexible team.

4 Results and discussion

When looking at the data during the EDA, we can see that a large amount of data values are missing, especially at the score columns. Column 13, User_Count, is missing more than 50% of the total values. The score columns can be dropped, but it could risk an underfitted model.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16719 entries, 0 to 16718
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Name             16717 non-null  object
 1   Platform         16719 non-null  object
 2   Year_of_Release  16450 non-null  float64
 3   Genre            16717 non-null  object
 4   Publisher        16665 non-null  object
 5   NA_Sales         16719 non-null  float64
 6   EU_Sales         16719 non-null  float64
 7   JP_Sales         16719 non-null  float64
 8   Other_Sales      16719 non-null  float64
 9   Global_Sales     16719 non-null  float64
 10  Critic_Score     8137 non-null   float64
 11  Critic_Count     8137 non-null   float64
 12  User_Score       10015 non-null  object
 13  User_Count       7590 non-null   float64
 14  Developer        10096 non-null  object
 15  Rating           9950 non-null   object
dtypes: float64(9), object(7)
memory usage: 2.0+ MB
```

Image 1: A summary of the data, showing missing values, using the data set.

What score values are missing, and why? Here we see that many games around the year 2010 are missing scores, and not many games in the years pre-2000 are missing scores. The years pre-2000 not missing many scores can however be due to there not being released as many games in that time period, as seen in the chart below.

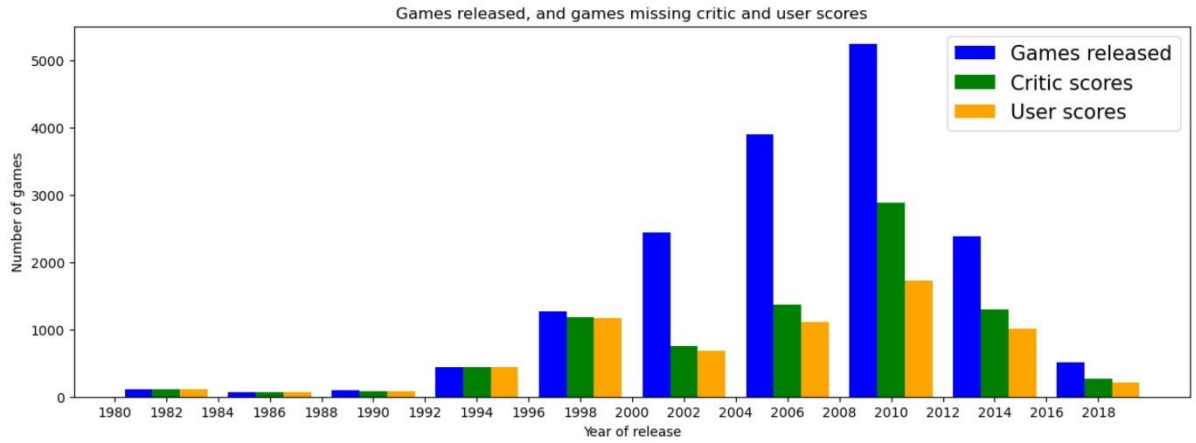

Image 2: A histogram showing games released alongside missing scores, using the data set.

It might be best to handle missing data before checking correlations between columns, but since there are a lot of missing values, there could be trust issues after the possible imputations.

Sales in regions seems to have more correlation with global sales, which is not surprising. These sales columns might be too similar/connected to Global_Sales and create overfitting if they are included in the model training.

Critic count seem to have some correlation, which seem logical. A popular game might get more attention from critics. Critic score and user count also have some correlation, but there aren't many strong correlations with global sales overall.

Year or release, platform, genre, publisher and developer seem to have weak correlations and could possibly be removed.

This leaves the model training with the columns that are missing a lot of values, which isn´t a promising sign.

	Year_of_Release	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Count	Platform_cat	Name_cat	Rating_cat	U
Year_of_Release	1.000000	-0.092562	0.003842	-0.168386	0.037700	-0.076433	0.011411	0.223407	0.175339	0.172032	-0.003648	0.174716	
NA_Sales	-0.092562	1.000000	0.765336	0.449598	0.638654	0.941010	0.240755	0.295413	0.246429	0.040758	0.012128	0.055807	
EU_Sales	0.003842	0.765336	1.000000	0.435068	0.722796	0.901239	0.220752	0.277533	0.283360	0.044896	0.006186	0.068188	
JP_Sales	-0.168386	0.449598	0.435068	1.000000	0.291096	0.612300	0.152593	0.180219	0.075638	-0.079609	0.016284	-0.090056	
Other_Sales	0.037700	0.638654	0.722796	0.291096	1.000000	0.749242	0.198554	0.251639	0.238982	0.054925	-0.006589	0.089971	
Global_Sales	-0.076433	0.941010	0.901239	0.612300	0.749242	1.000000	0.245471	0.303571	0.265012	0.026729	0.010845	0.044386	
Critic_Score	0.011411	0.240755	0.220752	0.152593	0.198554	0.245471	1.000000	0.425504	0.264376	0.018264	0.029438	0.040429	
Critic_Count	0.223407	0.295413	0.277533	0.180219	0.251639	0.303571	0.425504	1.000000	0.362334	0.230113	-0.000979	0.248871	
User_Count	0.175339	0.246429	0.283360	0.075638	0.238982	0.265012	0.264376	0.362334	1.000000	-0.006800	-0.038648	0.094063	
Platform_cat	0.172032	0.040758	0.044896	-0.079609	0.054925	0.026729	0.018264	0.230113	-0.006800	1.000000	0.014966	0.219699	
Name_cat	-0.003648	0.012128	0.006186	0.016284	-0.006589	0.010845	0.029438	-0.000979	-0.038648	0.014966	1.000000	-0.000337	
Rating_cat	0.174716	0.055807	0.068188	-0.090056	0.089971	0.044386	0.040429	0.248871	0.094063	0.219699	-0.000337	1.000000	
User_Score_cat	0.186572	0.067547	0.067710	-0.114055	0.090932	0.045728	0.307092	-0.005518	0.027021	0.094021	-0.015465	0.649548	
Genre_cat	-0.129007	0.019268	0.019145	0.030957	0.011820	0.023955	0.140810	-0.010218	0.008179	0.025860	-0.002360	-0.027279	
Publisher_cat	0.031729	0.002621	0.010575	0.051612	0.013072	0.016629	0.000688	0.036075	0.016235	-0.010843	0.027320	-0.006805	
Developer_cat	0.197920	0.073501	0.082371	-0.087047	0.099543	0.060033	0.022942	0.034344	-0.012620	0.117190	0.008928	0.541666	

Image 3: A correlation matrix, using the data set.

We decided to do two versions of the data going forward, one more simple where rows with missing values were removed, and one where transformations where made by filling missing values with mode and median.

With a linear regression model using the removed rows - technique, we got the following results.

Multiple linear regression	
R-squared	~0.16
RMSE	~1.89
Mean Absolute Error	~0.76
Mean Squared Error	~3.58

With a linear regression model using the transformation technique, we got the following results.

Multiple linear regression	
R-squared	~0.16
RMSE	~1.86
Mean Absolute Error	~0.56
Mean Squared Error	~3.46

The model trained on the transformed dataset was slightly better, but the difference was minimal, and none of the models were good.

We decided to train a new model, this time a decision tree model named CatBoostRegressor. There might be only one column needed for each node of the tree, which means that this model could provide a better result.

With the CatBoostRegressor model using the removed rows -technique, we got the following results.

Multiple linear regression	
R-squared	~0.47
RMSE	~2.14
Mean Absolute Error	~0.51
Mean Squared Error	~4.60

R-squared provided a much better result, while RMSE was a bit worse. This model is most likely the best so far.

The next step was to explain the output of this CatBoostRegressor model.

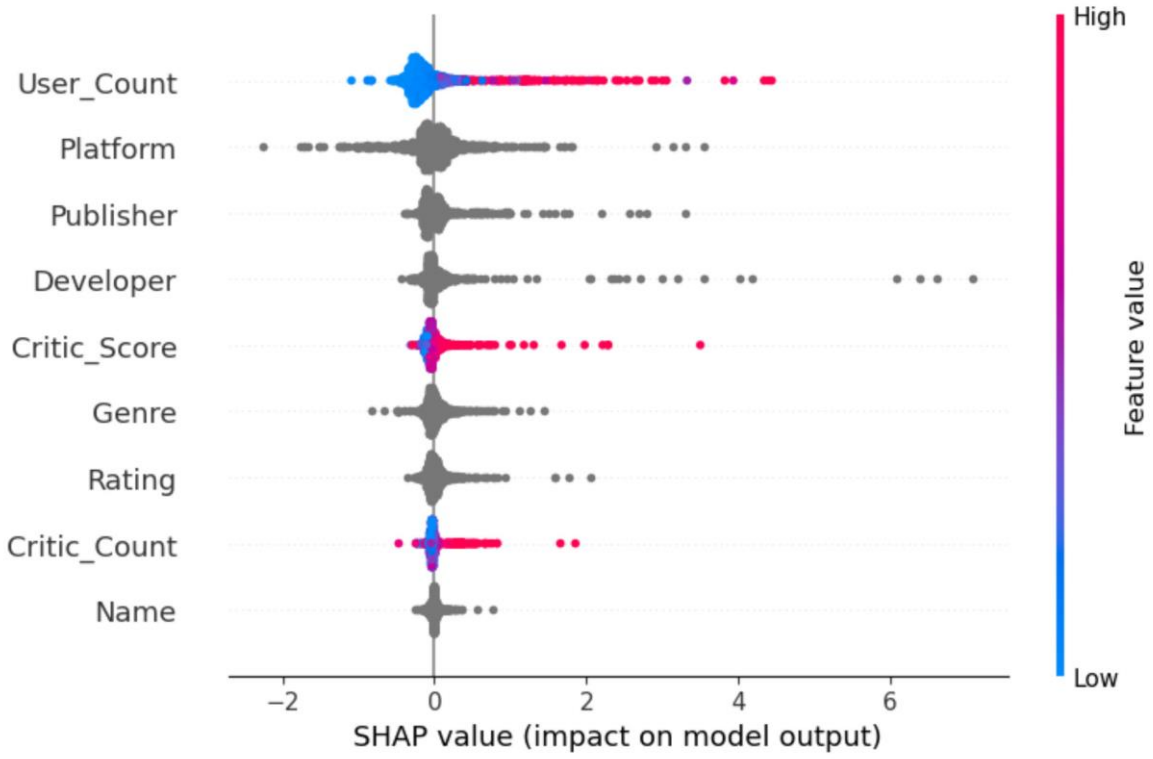

Image 4: A summary plot using SHAP values, the trained CatBoostRegressor model and the test set.

User_Count is the most important feature in this model. High User_Count counts (red) are at a higher shap value, which means higher counts tend to positively affect the output. This was somewhat visible in the correlation matrix in the EDA, although Critic_Score seemed to have a stronger correlation there.

Grey represents categorical values which cannot be scaled in high or low.

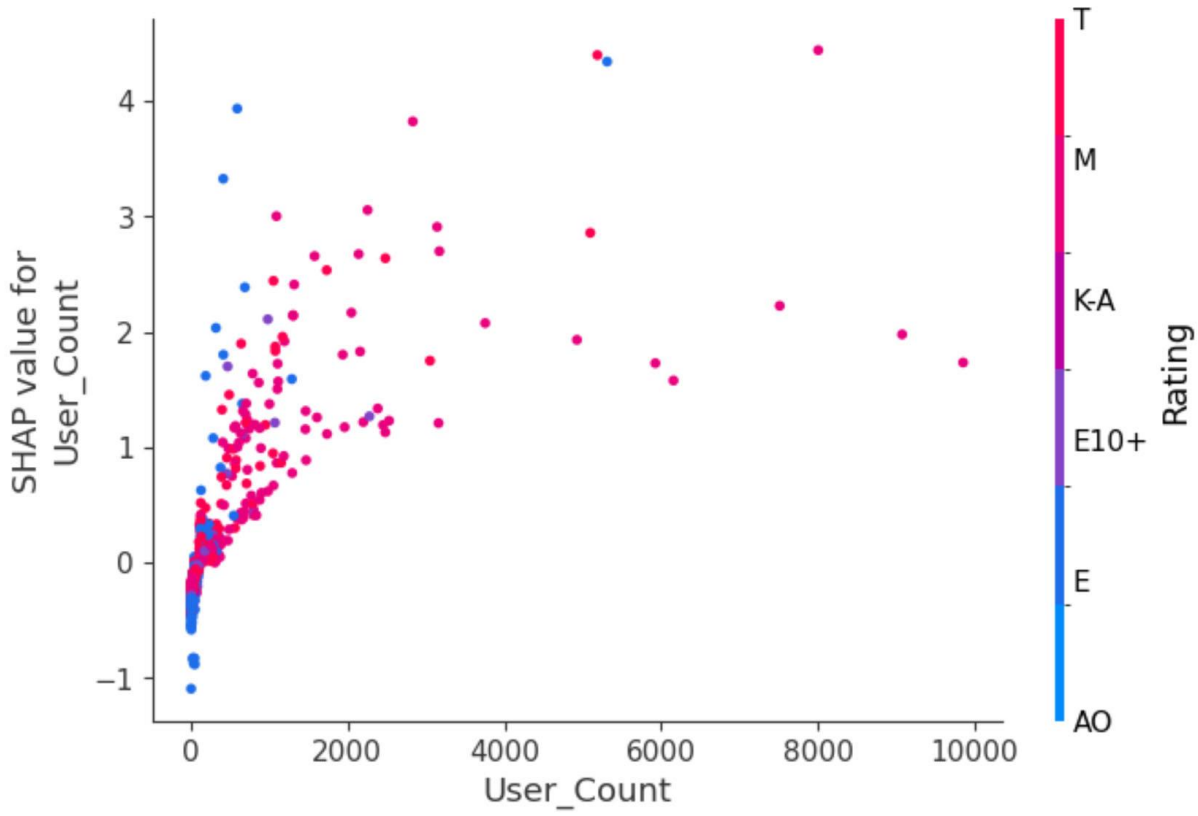

Image 5: A dependence plot for User_Count and feature, using SHAP values, the trained CatBoostRegressor model and the test set.

Datacamp explains that a dependence plot is a type of scatter plot that displays how a model's predictions are affected by a specific feature (Subscription Length). On average, subscription lengths have a mostly positive effect on the model (Datacamp, n.d).

Here we can see that it´s mostly Teen and Mature games with high User_Count counts (red) that are at a higher SHAP value, which means their higher counts tend to positively affect the output.

E -Everyone rating have a game that is low with a high SHAP value, which means lower User_Count counts tend to positively affect the output. This while having a high User_Count. It can also be viewed as an outlier.

Below we examined the first sample in the testing set to determine which features contributed to the "0" result. This was made with a force plot.

Image 6: A force plot for a specific instance, using SHAP values, the trained CatBoostRegressor model and the test set.

Positive SHAP values are displayed on the left side, and the negative on the right side. The highlighted value is the global sales prediction for that observation. We can see that the User_Count has lower SHAP value in this instance, and that it contributes negatively to the prediction. A higher Critic_Score has alone contributed positively.

With Power Bi, and the dataset with removed rows with values, we tried to highlight some of the features. Below is a scatter plot showing global sales and user count. Here we can see a positive trend on global sales as the user count increases. There´s also possible outliers shown.

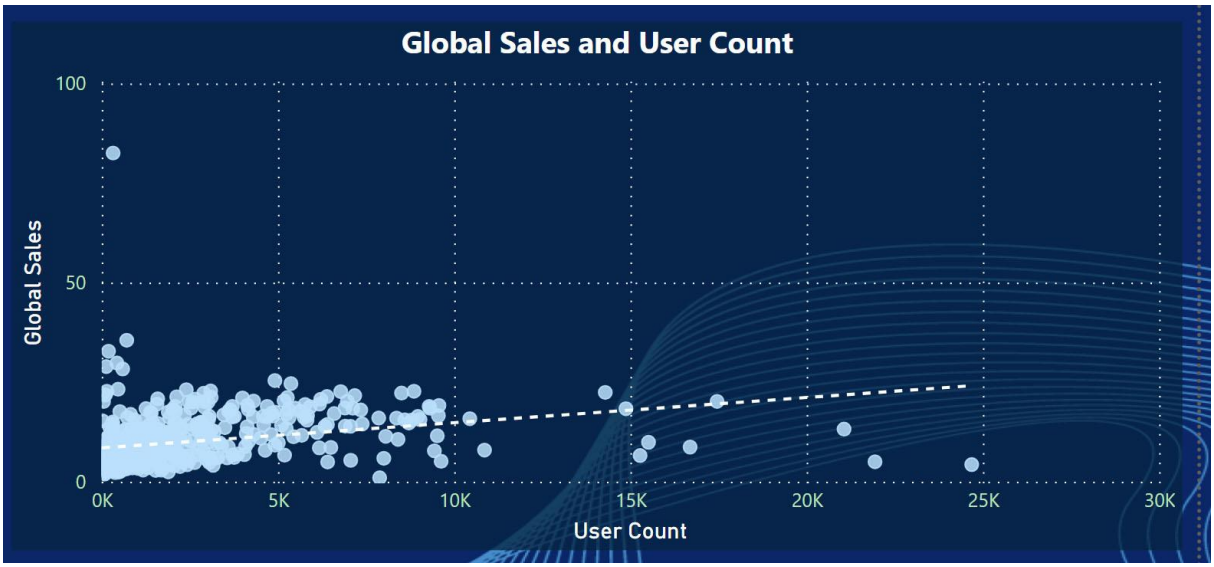

Image 7: A Power Bi scatter plot showing global sales and user count with a positive trend, using the removed rows dataset.

Below is a bar plot showing the top ten video game developers in terms of developer games count.

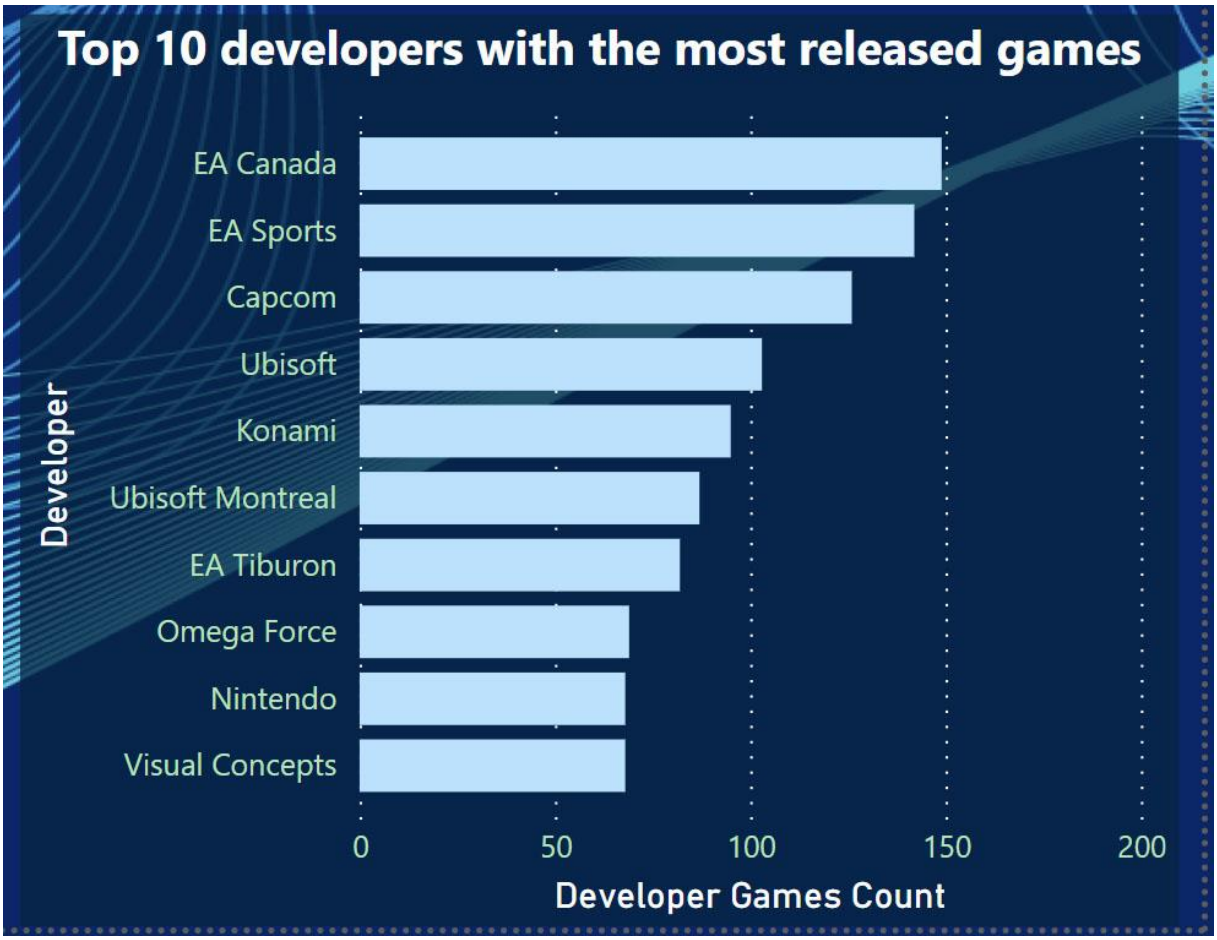

Image 8: A Power Bi bar graph showing the top ten game developers, using the removed rows dataset.

Below is a time series chart showing global sales over year of release, and even though data is removed, we can still see the same trend from “image 2”, with a maximum at somewhere around year 2008.

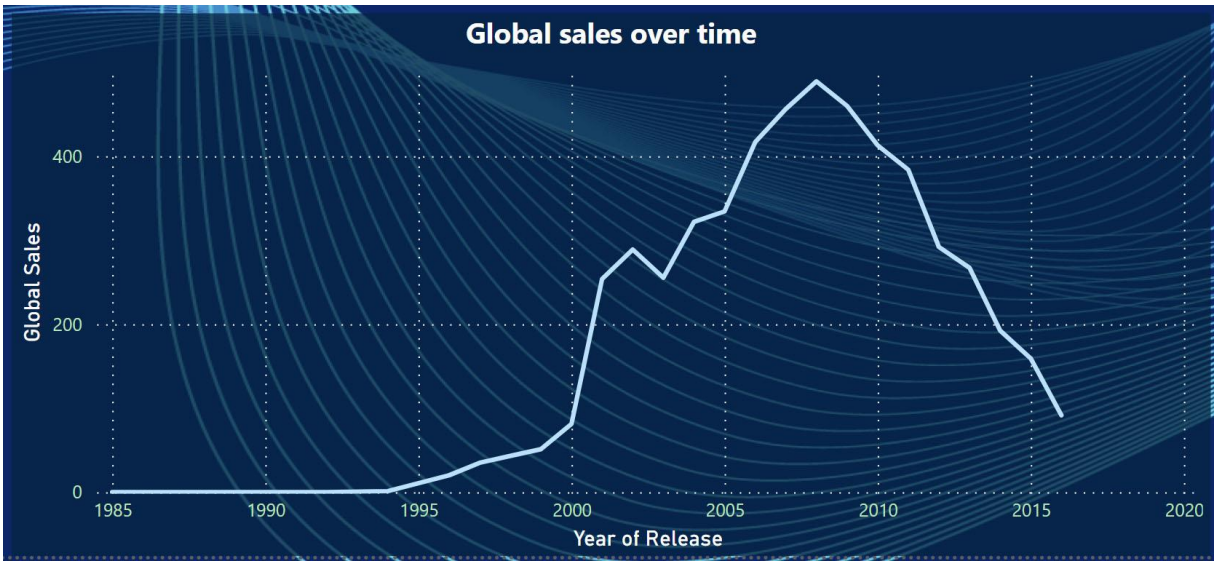

Image 9: A Power Bi time series chart showing global sales over year of release, using the removed rows dataset.

5 Conclusions

5.1 Is it possible to use this dataset to create a model predicting global sales?

The data was problematic. Lots of values were missing, which lead us to make a decision to, in the end, remove all rows with missing values. This is not ideal when training a model, and even though a model would be trained with good results, there is a risk of underfitting a model depending on all the data rows removed. If those rows/data points had values missing for a special reason, then the model would probably be unable to process them when encountering similar data points when doing predictions. User_Count and User_Score were some of the most important features, but also some of the features missing the most values.

The linear regression model results were poor. The decision tree model was better, but still not good enough. It might be possible to train a model predicting global sales by using this dataset, but it requires more work and testing. There are several possibilities for proceeding, such as selecting and trying more models and tuning hyper parameters using cross-validation. The best course of action would probably be to find the missing data, because “garbage in = garbage out”.

5.2 Can we find the feature(s) explaining the global sales (the output of the model)?

The CatBoostRegressor model, while not perfect, still contained some interesting information about the features used to train it. With the SHAP tool, we found that User_Count was the feature with the most impact on the model predictions. This is also one of the features missing the most values in the data set. A high User_Count made a positive impact on the global sales, which seem very logical. In power bi, we could also find a positive trend in a scatter plot when looking at User_Count and global sales.

The next most important feature was ‘Platform’, ‘Publisher’ and ‘Developer’. There is room for possible improvements here by looking at features that may correlate too much, and maybe join or remove any of them. A publisher like Nintendo will most likely release a game on its own platform, and the model training might benefit by joining all the Nintendo platforms ‘Wii, Switch…’ as a single Nintendo platform.

Some features were left out during training, to try out new combinations of the data the model got to train on. This could of course also be changed. This is a first cycle of this report, and I might revisit it in the future to make improvements.

6 Self evaluation

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

   Arbetet och grupparbetet har gått väldigt bra tycker iaf jag. Själva datan har varit svår att arbeta med då den kändes väldigt bristfällig, men med diskussioner i gruppen så kom vi fram till en gemensam plan framåt, att testa två olika versioner av datasetet, borttagna rader/transformerad data. Modell-träningen blev inte så lyckad ändå, men då kom vi fram till att vi skulle testa en ny modell. Den blev inte jättebra heller, men iaf bättre.
   Det var klurigt att förstå SHAP, men även där kom grupp-diskussion till hjälp, samt googlande efter förklaringar på webben.
   Den andra personen i min grupp var på semester den första veckan, men vi hade kontakt före kursen började där vi kom överens om vad vi skulle arbeta med för projekt. Under den första veckan började jag med EDA och några grafer, sedan kom gruppmedlemmen tillbaka och började genast bidra så att vi tillsammans lyckades slutföra projektet i tid.

2. Vilket betyg du anser att du skall ha och varför.

   Jag tycker att vi båda förtjänar ett VG då vi har jobbat effektivt tillsammans, vi har hjälpt varandra att förstå olika saker, löst saker i varandras kod, diskuterat och planerat/satt upp målstolpar med varandra, uppmuntrat varandra och kommit med feedback. Vi har kunnat komma med ideer och vi har kunnat justera kursen utan att någon inte hänger med. Själva projektet slutade inte i en perfekt tränad model, men under förutsättningarna så tyckte jag vi gjorde det bra. Dataanalysen har kritiskt granskats, och det har reflekterats över valen gällandes t.ex. data-behandlingen, modell-val, features o.s.v. Med hjälp av SHAP har vi kunnat fördjupa oss i vissa intressanta delar, och med power bi har vi kunnat visa att vi kan visa upp datan och relevant information även där.

3. Något du vill lyfta fram till Antonio?

   Det var en rolig kurs! Friheten har ökat mitt intresse av att läsa mer utanför kursmaterialet.

7 Appendix

Link to Jupyter Notebook:
https://github.com/dawnbanawn/ML_Game_Sales/blob/main/Video_Game_Sales_v7.ipynb

Link to Power Bi presentation:
https://github.com/dawnbanawn/ML_Game_Sales/blob/main/video_game_sales_v02.pbix

References

Datacamp. An Introduction to SHAP Values and Machine Learning Interpretability. Gathered 21 October, 2024, from Datacamp ´s website https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

Kaggle. Video game sales with ratings. Gathered 7 October, 2024, from Kaggle´s website https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings/data

Medium. MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? Gathered 19 October, 2024, from Medium´s website https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

Statista. Video game industry - Statistics & Facts. Gathered 8 October, 2024, from Statista´s website https://www.statista.com/topics/868/video-games/

Statology. Adjusted R squared interpretation. Gathered 18 October, 2024, from Statology´s website https://www.statology.org/adjusted-r-squared-interpretation/

Towards Science. CatBoost regression in 6 minutes. Gathered 16 October, 2024, from Towards Science ´s website https://towardsdatascience.com/catboost-regression-in-6-minutes-3487f3e5b329

Vitaflux. Decision Tree Regression vs Linear Regression: Differences. Gathered 18 October, 2024, from Vitaflux ´s website https://vitalflux.com/decision-tree-regression-linear-regression-differences-examples/