Videogames

What drives global sales



Daniel Hemgren

EC Utbildning DC 23

Projektkurs

20241101

# Abstract

The Video game industry have released around 16000 titles since the 80s and what drives the sales at a global level? That is the main question that we are trying to answer in this report. Two different modeling approaches are tried one parametric (Linear Regression) and one non-parametric (The gradient booster -Cat Boost). One Power BI visualization is also made.

**Skapas automatiskt i Word genom att gå till Referenser > Innehållsförteckning.**

Table of content

[Abstract 2](#_Toc181432120)

[1 Introduction 1](#_Toc181432121)

[1.1 Research questions 1](#_Toc181432122)

[2 Teori 2](#_Toc181432123)

[2.1 Linear regression 3](#_Toc181432124)

[2.2 Gradient Boosting 4](#_Toc181432125)

[3 Method 5](#_Toc181432126)

[3.1 Datacollection 5](#_Toc181432127)

[3.2 Preprocessing/EDA 5](#_Toc181432128)

[3.3 Modeling (research question 1) 6](#_Toc181432129)

[3.3.1 Linear regression 6](#_Toc181432130)

[3.3.2 Category booster 6](#_Toc181432131)

[3.4 Power Bi(research question 2) 6](#_Toc181432132)

[4 Results and discussion 8](#_Toc181432133)

[4.1 Research questions with answers 9](#_Toc181432134)

[5 Concluding remarks 10](#_Toc181432135)

[6 Självutvärdering 11](#_Toc181432136)

[7 References 12](#_Toc181432137)

# Introduction

The global sales of video games totals over 8 billion copies since the 'beginning' of the industry. The dataset used in this work contains the majority of this and starts in the 1980s. Sales take place across clearly distinct gaming platforms and game genres. Is it possible to see any factors that can explain why some games sell more and other games less.

(One thing to keep in mind throughout is that most VideoGame datasets are ’external data’, no data is from ’within the games’ like pixels, sounds, button combinations etc. that is probably hard to find in this industry.)

## Research questions

Which factors have been most important in explaining video game sales? That is the main question in this report. This question is materialised in the two research questions below:

1. Can we find the feature(s) explaining global sales?

2. Can any feature(s) explaining the global sales, be clearly presented in power bi?

# Teori

The broad theory behind this report is the theory of Statistical Learning, which focuses on finding the function (f) between different sets of data. Primarily, it provides tools for understanding data (Hastie, 2023, p. 1).

Statistical learning can be supervised or unsupervised. Supervised statistical learning broadly involves creating models to estimate an output based on one or more inputs. Unsupervised statistical learning, on the other hand, has inputs but no defined outputs; nonetheless, it allows us to learn about structures and relationships in the data.

The specific theory in this report revolves around the theory of linear regression modeling. This theory is well-suited for purposes where a high degree of "explainability" is important. Detailed understanding of how each part of the input affects the output can be achieved with these models, a process known as making inferences (Hastie, 2023).

Regression can also be used to model discrete outputs; in such cases, the models are called logistic regression and often use the logistic function.

The main focus of this report is on Linear Regression, a supervised method for creating models between inputs and a continuous output. This is also referred to as "creating a linear model between an output with a continuous label and one or more input features." Linear regression may seem "too simple," but it is highly powerful and has significant advantages in interpretation and explanation. Linear regression prevents the model from becoming a "black box" (https://en.wikipedia.org/wiki/Black\_box, n.d.).

Within supervised learning:

1. Parametric Models

Parametric models make specific assumptions about the functional form of the relationship between input features and output targets. They typically rely on a finite set of parameters that define the model.

2. Non-Parametric Models

Non-parametric models do not make strong assumptions about the functional form of the relationship between input features and output targets. Instead, they adapt to the data and can capture complex relationships without being constrained by a predefined structure.

Decision Trees: They do not assume any specific form of the relationship between features and target values. They partition the feature space into regions based on feature values, creating a tree structure that can handle both regression and classification tasks.

Random Forests and Gradient Boosting: These are ensembles of decision trees and inherit their non-parametric nature, allowing them to model complex relationships.

Category booster represents somewhat of an opposite to linear regression (no mathematical assumptions like in the ’General linear model’).

## Linear regression

Linear regression is a statistical method used to model the linear relationship between one or more independent variables (features) and a continuous dependent variable (target). The goal of linear regression is to create a linear equation that describes the expected average effect of the independent variables on the dependent variable. Linear regression relies on several assumptions, including a linear relationship between variables, normal distribution of residuals, homoscedasticity (constant variance of residuals), and independent observations (Hastie, 2023).

Calculating the least squares sum is the most common method to fit this model to the data (Hastie, 2023, p. 59).

In its most basic form, simple linear regression appears as shown below, where

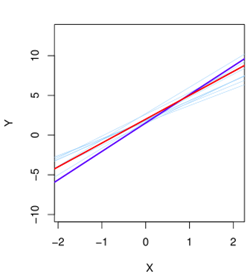


Y is "approximately modeled" by the two beta terms. The first term is the intercept, and the second is the slope of the line (Hastie, 2023, p. 61).

The true population relationship (which we cannot observe) is shown below:



With multiple different samples, the regressions would look like the blue lines below, but their average would lie close to the unobservable population regression, shown in red below.



At a 95% significance level, the fixed but unknown Beta must lie within 2 standard errors from the "β-hat" according to the interval for "β-hat" below. (In the purpose section of this report, 99% significance was chosen, which corresponds to 3 standard errors.)



The multiple regression model is an extension of simple regression, where each Beta has its own slope in relation to the dependent variable. The fixed but unknown Beta coefficients are estimated similarly to in simple regression, where a chosen number of standard errors around the estimates provides an interval for Beta at a given significance level. ±2 standard errors indicate 95% significance, and ±3 standard errors, for example, indicate 99% significance. In this model, it is essential to interpret Beta1 as the average effect on Y of a one-unit change in X1, holding all other predictors constant.



Here too, the residuals must meet the requirements of being normally distributed, homoscedastic (constant variance), and independent of each other (not autocorrelated).

Multiple regression has an inherent risk: with too many inputs (sometimes more than observations), some significance will always be achieved by pure chance (Hastie, 2023, p. 78). Therefore, tools such as the F statistic are used to adjust for the number of predictors and avoid this problem.

Criteria like Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or adjusted R-squared are used to penalize model complexity and select the best model. Alternatively, the data can be split into training and validation sets multiple times, models fitted on the training sets, and their performance evaluated on the validation sets. The model with the best average performance across validation sets is chosen. This requires computational power and is known as resampling techniques (Hastie, 2023) p.197

## Gradient Boosting

Is in the decision tree paradigm and is well described here:

”Gradient Boosting works by sequentially adding predictors to an ensemble, each one correcting its predecessor. However, instead of tweaking the instance weights at every iteration this method tries to fit the new predictor to the residualerrors made by the previous predictor”. (Hands-On\_Machine\_Learning\_with\_Scikit-Learn-Keras-and-TensorFlow-2nd-Edition-Aurelien-Geron)

Where an ensemble is any method that combines several weak learners into a strong learner, a weak learner is slightly better than random guessing. (Kearns(1988) & Thoughts on Hypothesis Boosting)

Gradient boosting is essentially a process of constructing an ensemble predictor by performing gradient descent in a functional space. (Vorobev, 2017)

Cat boost is a special form of gradient booster

CatBoost is a high-performance gradient boosting algorithm designed to handle categorical data efficiently without extensive preprocessing. Introduced by (Vorobev, 2017) it addresses common issues like prediction bias by using a novel ordered boosting approach, which minimizes target leakage. Additionally, CatBoost's efficient handling of categorical features and compatibility with both CPUs and GPUs make it suitable for large-scale, complex data tasks.

SHAP (SHapley Additive exPlanations): CatBoost supports SHAP values, which provide a way to quantify the contribution of each feature to individual predictions, offering more granular insight into model decisions.

# Method

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.

## Datacollection

The data comes from Kaggle: https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings/data.

The dataset consists of 16 columns and around 16,000 game titles. Some columns have certain gaps, and considering the research questions, columns that are not useful or relevant might be dropped.

## Preprocessing/EDA

The research questions aim to explain global sales, so sales figures for specific geographic regions are dropped since they are subcomponents of global sales, which is also evident in the correlation matrix.

# 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 arent many strong correlations with global sales.

# 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 is a bad sign.

Initial cleaning for redundancy and irrelevance: : 'User\_Score', 'NA\_Sales', 'JP\_Sales', 'Other\_Sales', 'Year\_of\_Release', 'EU\_Sales' These all show correlation close to 0 with our target , or are terms making up our target (for example JP\_sales are part of global\_sales)The data is mixed with categorical an numerical columns and some missing values. This can be handled in several ways. Deleting and imputing values are the two main ways. Deleting missing values is the first approach , and then to be compared to imputing values. When imputing values it is important to impute separately on the train and test set so as to avoid ’data leakage’ into the test set.

## Modeling (research question 1)

Global sales is to be modeled according to research question1 so that is the target.

### Linear regression

Rows with empty values are deleted to create the first ’baseline model’. A Linear model is trained after splitting into train and test sets. The test set is then used to predict our target with the trained model. The result shows that this first baseline model is not explaining the target well.

The next approach is to try and impute values into empty cells instead of deleting rows. The median is imputed into the numerical columns and the mode into categorical columns. This is done separately in the train and test set so as to not spill info from the train set into the test set (data leakage) . The trained model does not show much improvement when used for prediction and compared to the target variable in the test set. To try and tweak something out of linear regression in this situation seems hard, a new approach is needed if a useful model is to be developed and the research question 1 is to be answered.

### Category booster

The main idea is to search for some model that will be somewhat of an opposite to linear regression, that a category booster is non parametric is a good first step.

Cat boost can also work directly with no encoding of category columns, it should model the data different to a Linear regression and is therefore chosen. The removed rows dataset is used for simplicity. No real difference between the sets in the linear regression modeling but it could teoretically mean something here. After training the Catboostmodel the model is used for predictions. The result is an improvement(MAE, R2) on the Linear regression, this will be elaborated on further in the results section. The SHAP visuals show that a large part of most features are not affecting our target by much (the fat part of the violin plots are mostly close to 0 in SHAP value). More about that in the results section.

## Power Bi(research question 2)

The dataset is visualized in Power Bi so that the features can be explored against our target, Global sales. Here it shows that even for the best predictors like User Count the relationship with Global sales is not a good linear relationship, although at higher User Counts the relationship looks barely linear to the naked eye.

# Results and discussion

|  |  |
| --- | --- |
| **MAE for different models** | |
| Linear Regression (removed rows) | 0,76 |
| Linear Regression (imputed rows) | 0,67 |
| Cat Boost Regressor | 0,52 |

Table 1: Mean average error (MAE) for the three models.

The MAE is lower for Cat Boost Regressor showing that it works better at least in the middle of the distribution.

|  |  |
| --- | --- |
| **RMSE for different models** | |
| Linear Regression (removed rows) | 1,88 |
| Linear Regression (imputed rows) | 2,29 |
| Cat Boost Regressor | 2,16 |

Table 2: Root mean square error (RMSE) for the three models.

Not big difference in result shows that outliers play large part in the modeling.

|  |  |
| --- | --- |
| **R2 for different models** | |
| Linear Regression (removed rows) | 0,17 |
| Linear Regression (imputed rows) | 0,14 |
| Cat Boost Regressor | 0,46 |

Table 3: R2 for the three models.

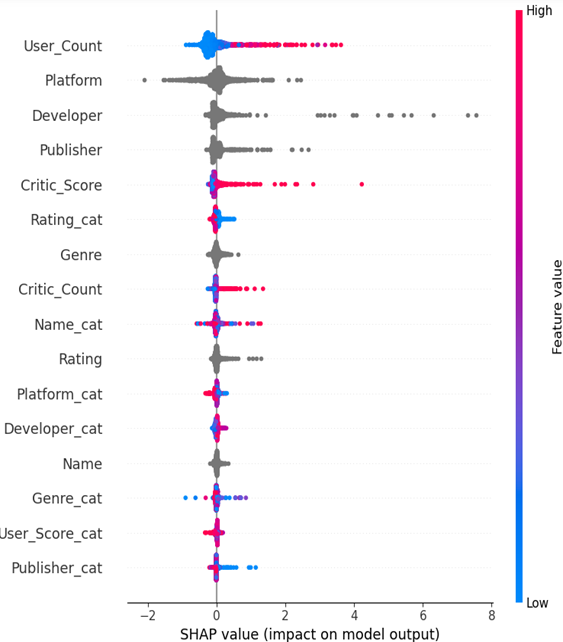


Table 4: SHAP values for Cat Boost regressor model

The SHAP visuals show that a large part of most features are not affecting our target by much (the fat part of the violin plots are mostly close to 0 in SHAP value). User count is the feature that affects Global sales(model output) most clearly. Critic score also affects Global sales.

## Research questions with answers

1. Can we find the feature(s) explaining global sales?

User count and Critic score are both reasonable but weak explainers for Global sales.

2. Can any feature(s) explaining the global sales, be clearly presented in power bi?

User Count is visualised against Global sales in Power BI. (the Power BI report can take any other feature as well)

# Concluding remarks

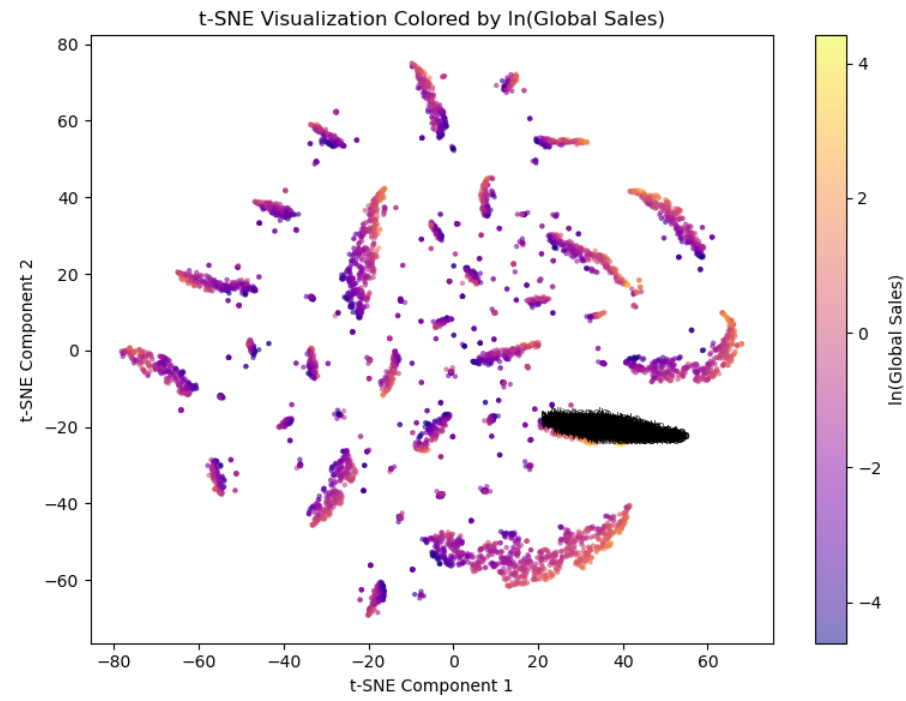
The results section shows that some outliers play a heavy part in the prediction residuals. Cat Boost is clearly the best model ’in middle of target distribution’ but when looking at RMSE some big outliers make the predictive ability of the catboost model less than ideal also.

Another approach might have helped: To create model in just the middle of target distribution. MAE shows that catboost improves predictions over Linear Regressions at least in the middle of the distribution.

To predict or make inferental statements about global sales clearly takes another approach. Another dataset is probably needed to make this a really good model. Data from ’inside’ the games (like play time/button combinations/graphics data/sound data etc .would be interesting to model but is probably really hard to find)

’All models are wrong but some are useful’ -- this became one of the models that were not useful I think.

If I were to model this dataset again I would explore the dataset a lot deeper and make that the main research question with dimension reduction techniques, like t-SNE. I experimented with that but didnt find any obvious ways to approach the modeling so we settled on the ’ortodox ways’ with data preprocessing. Here are game publishers in dimension reduction view(t-SNE) with Nintendo annotated and games coloured by logaritmic(Global Sales). Although Nintendo cant be seen easily it becomes ’black blob’, It shows that publishers are clearly separated at least.



# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.   
   Samarbetet med Dan flöt på riktigt bra och inga direkta utmaningar i arbetsprocessen, vi beslutade oss för tydliga frågeställningar tidigt och sedan höll oss till den strukturen i vårt Agila samarbete. Valen vad vi gör med datasetet blev de stora besluten.
2. Vilket betyg du anser att du skall ha och varför.   
   På hur bra modellen blev inte VG , på förståelsen av vad en annan approach kunde gett och lärdomar kanske VG.
3. Något du vill lyfta fram till Antonio?

’Proffs fokuserar på datat och amatörer på modellen’ ..eller typ så.. tror jag du nämnde på en lektion.. jag tycker jag lärde mig det i detta arbete vad det innebär. I detta fall kunde ambitionsnivån ändrats till att undersöka datasetet på djupet och mer skapa ’proof of concept’ för någon annan att modellera på.

# References

*Hands-On\_Machine\_Learning\_with\_Scikit-Learn-Keras-and-TensorFlow-2nd-Edition-Aurelien-Geron.* (u.d.).

Hastie, T. (2023). An Introduction to ststistical learning 2nd.

Kearns and Valiant (1988, 1. (u.d.).

Kearns(1988), M., & Thoughts on Hypothesis Boosting, U. m. (u.d.). *Kearns and Valiant (1988, 1989).*

Vorobev, A. (2017). CatBoost: unbiased boosting with categorical features. *arxiv.org*.