Continual Assessment (2025)

Predict Steam Game Prices

Robert Smyth

Computing with Games Development, MTU Kerry

Abstract

This project explores if it is possible to predict the price of a Steam Game. I will use a dataset of over seventy thousand games each with a set of values regarding that game. Following the CRISP-DM process , I will try to estimate a games price, from information like estimated owners, release year, number of supported languages and more. Most of the work I did went into preparing the dataset 'games.csv', as retrieved from Kaggle as of November 2025, as many columns in this dataset contained messy text, line breaks, extra spacing, and missing data. I then cleaned these fields and converted some of them into numeric values where I could then manipulate them. I removed unusable fields and used Scikit-Learn to train and test a Linear Regression Model.

I found the model performed poorly, with an R2 score of 0.05 and high MAE and RMSE, the model underestimated expensive games and overestimated cheap games. I believe this indicates that linear regression cannot predict a games price. Future work could explore more advanced models like Random Forest and include more features such as Reviews to improve predictions.

Keywords: CRISP-DM; Regression; Steam; Games; Price; Predict;

1. Introduction

Deciding a Price for a Video Game can depend on many features, for example how popular the publisher is, when it is released, how much content it offers and even how many languages it can support. For Indie Developers, choosing a price can be very confusing, trying to match other Indie Games and not pricing the same as a AAA Studio. For this project, I had to follow the CRISP-DM steps. The idea is to check whether the data available regarding a list of Steam Games can be used to predict a games price.

The dataset I used had over 71,000 games and 39 different columns. Before I could model anything, a lot of work had to be done to deal with messy and missing data such as long text fields, mixed languages, special characters and such. This meant I had to start by figuring out how to clean everything and shape it into something I can use.

I wanted to see if I could predict the price of a Steam Game using only the numeric features in the dataset. To investigate this, I had to prepare the dataset, select usable numeric features and train a linear regression model. m

2. Background/Literature Review

In order to find similar research papers, I had to searched using Google Scholar. I searched through the various amounts of journals, articles and conference papers. I was looking for papers that used game data and numeric values to predict something about a game. Most of the research I found was focusing on sales and popularity of games rather than pricing. The studies I found gave me ideas on how game data can be used to predict something (Cunningham & Jane, 2021) (Huang, 2023) (Wenxiang Hu, 2024)

2.1. Game Sale

One paper I looked at was on predicting video game sales, using a dataset of over 16000 titles *(Wu, 2024)*. They explored different factors like ratings and publisher details and how they might relate to sales. They discussed and reported how advanced approaches can handle this type of data better than basic regression models, even though regression models can still reveal some trends.

Another paper examined how user and critic reviews can impact game sales (Huang, 2023)Their findings showed that reviews can influence how players make purchasing decisions, but that its not always straightforward. Although I focus on predicting the price, this showed how game related attributes can have an effect on some predictive outcomes.

2.2. Underfitting and Limits

An article on Underestimation Bias and Underfitting in Machine Learning published on Arrow for TU Dublin *(Cunningham & Jane, 2021)*explored how models could fail to capture patterns in data, hence resulting in Underfitting. Therefore, when a model relies on only a few features , such a numeric, as opposed to a large list of different pieces of information of a game, it can struggle. This makes them susceptible to missing data and directly relevant to the challenges I faced in my project.

2.3. Game Popularity

Another study looked at how different details of a game can affect game popularity on Steam *(Wenxiang Hu, 2024)*.They used details such as games price, tags and reviews to see which games would become more popular. They showed how it performed better once the dataset went through proper cleaning and processing.

Even though this paper looked at popularity rather than price, it supported the idea that Steam Game data can be used to build a predictive system.

3. Methods

I followed the CRSIP-DM methodology to try and predict Steam Game prices. Predicting game prices can help small developers decide how to price new games they are selling on Steam. I focused on numerical information such as estimated owners, release year, number of DLCs, concurrent users, required age and number of supported languages.

I used a dataset called ‘games.csv’ from Kaggle retrieved in November 2025, which has over 70,000 games and 39 columns of different information about each game. These columns included numeric, boolean and text data. I inspected it to find any missing data. A few columns did have missing data and these needed cleaning. Other columns like text descriptions and reviews were not used.

Next, I cleaned and transformed the data. I converted the release date to a single numerical value (year), I converted the estimated owners from a range to average value, I converted supported languages from a text field to a number via count, and I dropped unnecessary columns that were text heavy and not useful for my prediction. I also handled missing values by inserting calculated values for release year and estimated owners. I split the dataset into training and testing sets, 80/20, 80% for training and 20% for testing. I then checked that all the features were numeric.

I used Scikit-Learn and trained the model on the training set and tested it on the test set. Predictions were made on the test data, and a small table was created to compare actual vs predicted prices for the first 10 games in the dataset, Table.1.

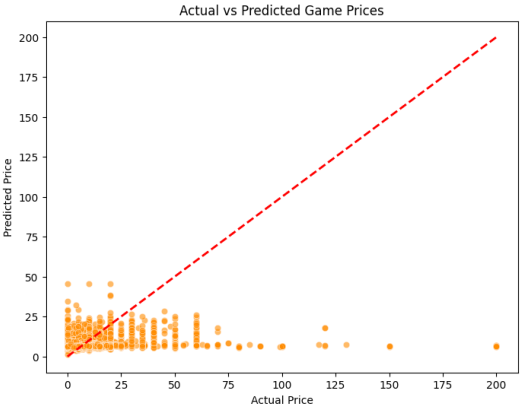
To check how well the regression model performed, I used the R2 score to see how much of the variation in price is explained by the model, the MAE to showed the average difference between the predicted and actual prices and the RMSE to show the size of errors in the predictions. I made scatter plots of actual and predicted prices to see how well the model performed. Most of the points were near the lower price range, resulting in high-priced games being underestimated and free or cheap games being overestimated.

4. Results/Findings

The Linear Regression model was trained on 80% of the dataset and tested on 20%. The predicted prices were compared with the actual prices.

Game Number	Actual Price($)	Predicted Price($)
1	9.99	6.07
2	0.00	5.95
3	0.00	6.62
4	29.99	6.69
5	11.99	6.56
6	12.99	6.76
7	11.39	13.21
8	14.99	12.61
9	4.01	6.78
10	1.99	6.23

This table shows that the model tends to underestimate the more expensive games and overestimate free or cheap games, this aligns with the low R2 score, observed as 0.05.



The following scatter plot shows that most predicted values cluster around the lower left, the lower prices. The red dotted line represents a perfect prediction, therefore points above the line are underpredicted and the points under the line overpredicted.

The predictions cluster around the mid-range prices due to underfitting, and numeric data didn't capture all the columns that can affect game prices like popularity, reviews or genre. Linear Regression would then give a rough approximate to predicting prices.

5. Discussion

The results from the linear regression model using CRISP-DM really shows that predicting a Steam Games price using only the numeric values on that game is actually very difficult. I believe this proves that simple linear regression isn't enough to make accurate price predictions. This was what I expected in the back of my mind, because I knew game prices depend on different factors that aren't in the dataset, but seeing the numbers shows how limited this model was.

An issue I had in this project was actually cleaning all the data, and transforming some of the columns to data I could use. This dataset had 39 pieces of information per game, and had missing values, strange formatting, multi-line text fields and more. Loading the CSV initially caused errors that need adjustments.

This however, was not just affecting me, other researchers running Steam datasets seemed to go through the same struggles. According to (Wenxiang Hu, 2024)working with big steam datasets always involves a lot of cleaning, and some missing entries can ruin predictions. (Huang, 2023)also pointed out that datasets with game metrics like reviews are rarely straightforward and always need careful preprocessing. Another struggle was the amount of data I could use. I tried using only numerical columns because that's what I knew how to handle. I can see that by not using some of the features available in the dataset left the model missing a lot of information. For example, a game could be very cheap but have high review scores, and an indie game could be expensive with low review scores, and none of that shows up in the numerical columns I used.

Therefore, Linear Regression is too simple to predict game prices reliably with only numerical features. The low R2 score, high MAE and RMSE, and the way the predictions cluster around the mid-range prices (Figure 1) all point to underfitting, resulting in the model needing more training time, and/or more features. As noted by Cunningham and Delany:

"The model underfits the data, failing to capture the variation in game prices and making errors even on the training set." (Cunningham & Jane, 2021)

6. Conclusion

In conclusion, this project shows that predicting Steam Game prices using only numeric features with Linear Regression is difficult. The low R2 score of 0.05 and high MAE/RMSE show that the model does not capture good enough information. Data cleaning and transformation are a critical part of this, however, excluding heavy text features like Reviews limited the predictability of the model.

Future improvements could involve using more advanced models such as Random Forest and including more features that can better capture the games' context and popularity and hence, a more accurate price prediction.

Bibliography

Cunningham, P. & Jane, D. (2021) Underestimation Bias and Underfitting in Machine Learning. Santiago de Compostela: School of Computer Science, TU Dublin. Available at: https://arrow.tudublin.ie (Accessed: 5 December 2025).

Huang, W. (2023) Research on the Prediction on the Sales of Electronic Games. HBEM – Highlights in Business, Economics and Management. Available at: https://drpress.org/ojs/index.php/HBEM/article/view/13594 (Accessed: 5 December 2025).

Wenxiang Hu, Y., Wang, R. & Xu, X. (2024) Machine Learning-Based Steam Platform Game Popularity. In: Proceedings of the 16th International Conference on Agents and Artificial Intelligence. Available at: https://www.scitepress.org/Papers/2024/128538/128538.pdf (Accessed: 5 December 2025).

Wu, K. (2024) Machine Learning Models-Based Video Game Sales Prediction. In: Proceedings of the 16th International Conference on Agents and Artificial Intelligence. Available at: https://www.scitepress.org/Papers/2024/132058/132058.pdf (Accessed: 5 December 2025).